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

The Divergence Hypothesis: Unmasking Lexical Interference and Label Bias in Mental Health NLP
Moustafa Hassan 1

Towards Unified Factuality Evaluation for Biomedical QA and Summarization: Aligning Metrics with Clinical Use-Cases
Mahule Roy and Subhas Roy 15

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 20

Analyzing Prompt Design Choices in Biomedical Information Extraction for Low-Resource Languages
Ayesha Khatun, Kadir Bulut Ozler, Steven Bethard and Egoitz Laparra 31

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 45

Agentic Feature Selection via LLM for Epileptic Seizure Detection
Aizierjiang Aiersilan and Xiaodong Qu 64

Training Biomedical Retrievers From Large-Scale Citation Contexts
Xing David Wang, Duy Le Thanh and Ulf Leser 75

Reliable Automated Triage in Spanish Clinical Notes: A Hybrid Framework for Risk-Aware HIV Suspicion Identification
Rodrigo Morales-Sánchez, Soto Montalvo and Raquel Martínez 84

Gold Label Errors in the SciFact Benchmark: An LLM-Assisted Annotation Audit
Julien Sylvestre 97

BioRAG: A Systematic Ablation Study of Retrieval Strategies for Biomedical Question Answering
Krushil Bhojani, Mayank Waghmare and Hima Bindu Nandyala 104

Post Hoc Agentic Refinement for Improving Precision in Multilingual Clinical Text De-identification
Justin Xu, Alistair Johnson, Thomas Lin, David Eyre and Rodolfo Quispe 115

Do Syntactic Features Help Biomedical Relation Extraction? An Empirical Study of Verb Token and Dependency Graph Augmentation
Mustafa Sikder and Ernest Kwegyir-Afful 128

Beyond Knowledge Graphs: PubMedBERT Embeddings as a Competitive Standalone Modality for Drug Re-purposing
Rishik Kondadadi and John E. Ortega 135

When Demographic Sensitivity Isn't What It Seems: Baseline-Aware Counterfactual Audits for Clinical NLP
Hyunwoo Yoo 141

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 156

Uncertainty-Aware Multi-Label Routing of Clinical Text to Surveillance Pathways
Agathe Zecevic, Sebastian Zeki and Angus Roberts 181

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 191

Effects of Adaptive Pretraining in Specialized Domains for Named Entity Recognition
Jack Lynam and Sam Henry 199

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 209

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 235

Diagnosing Lower Extremity Arteriovenous Diseases Using Agentic LLMs
Zicen Liao, Yunhao Sun and Matthew Purver 250

KGRxn-LLM: Knowledge Graph Enhanced Large Language Models for Molecular Reaction Reasoning
Weichen Liu, Qiyao Xue, Yuyang Wu, Olexandr Isayev and Natasa Miskov-Zivanov 268

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 282

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 292

Treating Decoder-Only LLMs as Encoders: A Simple and Effective Fine-tuning Approach for Named Entity Recognition
Ken Yano and Hiroya Takamura 312

A Multi-View Framework for Cross-Domain Nutrition Misinformation Detection in Social Media
Vishwaa Shah, Indika Kahanda, Andrea Arikawa, Asal Abbaszadeh and Richard Loftis 326

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 342

Systematic Evaluation of the Quality of Synthetic Clinical Notes Rephrased by LLMs at Million-Note Scale
Jinghui Liu, Sarvesh Soni and Anthony Nguyen 353

Bridging the Version Gap: Multi-version Training Improves ICD Code Prediction, Especially for Rare Codes
Jinghui Liu and Anthony Nguyen 372

EmCellLLM: Human Peri-Implantation Embryonic Cell Annotation Based on Large Language Models
Xiaorui Guo, Zhiwei Liu, Qianqian Xie and Sophia Ananiadou 382

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 392

Citation-Aware Continual Pre-Training for Biomedical Language Models
Masaki Asada, Tomoki Tsujimura, Tatsuya Ishigaki, Shusaku Egami, Ken Fukuda and Hiroya Takamura 407

TrackList: Tracing Back Query Linguistic Diversity for Head and Tail Medical Knowledge in Open Large Language Models
Ioana Buhnila, Aman Sinha and Mathieu Constant 413

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 426

PrionNER: A Named Entity Recognition Dataset for Prion Disease Biomedical Literature
An Dao, Nhan Ly, Thao Tran, Yuji Matsumoto and Akiko Aizawa 435

Evaluation of Multilingual Text Simplification for the Mental Health Domain: Exploring Small Language Models
Olga Pelloni, Sandra Just and Lars Bongo 464

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 475

Reading Between the Lines: Toward Translating Verbose Patient-authored Messages into Clinician-Formulated Questions
Sarvesh Soni, Madeline Bittner and Dina Demner-Fushman 481

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 490

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 502

When Does Retrieval Beat Direct LLM Diagnosis in Rare Disease? An Empirical Study of Ontology Coverage
Mohamed Elmofty and Ulf Leser 508

BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs
Nourah Salem, Elizabeth White, Michael Bada and Lawrence Hunter 519

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 531

BioConflict: A Benchmark for Evaluating Large Language Models in Biomedical Contradiction Detection and Consensus Synthesis
Ashwin Kirubakaran and Henry Gagnier 552

Tokenization Granularity and Medical Term Representations in Language Models
Vojtech Lanz and Pavel Pecina 559

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 572

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 595

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 605

MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP
Santiago Martinez Novoa, Lina Gomez Mesa, Juan Prieto and Ruben Manrique 617

When Evidence Conflicts: Uncertainty and Order Effects in Retrieval-Augmented Biomedical Question Answering
Yikun Han, Mengfei Lan and Halil Kilicoglu 630

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 644

Clinical Evidence and Patient Reviews: A Linked Dataset for Antidepressant Side Effects
Steven Au 656

A Deterministic Multi-Stage Retrieval Pipeline for Longitudinal EHR Question Answering
Shubham Agarwal, Thomas Searle, Richard Dobson, Ninoslav Majkic and Niko Moller-Grell 665

Interpretable ICD Code Classification with Faithful Sentence Extraction
Yichen Wang, Lian Hong, Masato Mizogaki, Shunnosuke Umeda, Toshimune Kenmotsu, Akihiro Tamura and Daniel Andrade 679

Evaluating LLM-as-a-Judge for Medical Term Simplification
Ioana Buhnila, Aman Sinha, Rohit Agarwal, Dilip K. Prasad and Mathieu Constant 687

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 695

Small LLMs for Biomedical Claim Verification: Cost-Effective Fine-Tuning, Structural Dataset Shortcuts, and Cross-Domain Generalization
Gaurav Kumar 704

Diagnosable ColBERT: Debugging Late-Interaction Retrieval Models Using a Learned Latent Space as Reference
François Remy 713

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 718

CENT: Context Engineering Framework for Normalization of Clinical Trial Procedures
Sanya Taneja, Ziqing Ji, Hans Verstraete and Ali Samadani 729

Agentic AI Architectures for SOAP Note Generation
Keno Hanken 742

VERICITE: Evaluating Sentence-Level Citation Faithfulness in Retrieval-Augmented Medical Question Answering
Yixian Ma, Bohao Chu and Norbert Fuhr 753

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 760

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 . 770

AAbAAC: An Annotated Corpus for Autoimmunity Information Extraction
Fabien Maury, Solène Grosdidier, Maud De Dieuleveult and Adrien Coulet 791

What Do Biomedical NER and Entity Linking Benchmarks Measure? A Corpus-Centric Diagnostic Framework
Robert Leaman, Rezarta Islamaj and Zhiyong Lu 801

Towards Grounded Hallucination Definitions for Biomedical Question Answering with Reproducible Examples from ClinIQLink
Brandon Colelough, Davis Bartels, Madeline Bittner and Dina Demner-Fushman 812

Can NLP Models Detect When One Publication Outweighs Twenty? Predicting Systematic Review Conclusion Changes
Ebrahim Alharbi and Mark Stevenson 843

VaxScope: Document-Level Structured Evidence Extraction from Immunization Systematic Reviews
Bahar Ilgen, Ebenezer Awotoro and Georges Hattab 853

Medical Context Variation: A source of impairment for Event classification
Aman Sinha, Marianne Clausel, Mathieu Constant and Xavier Coubez 864

KALIMBA: Knowledge-Assisted Literature Mining for Biological Interaction Analysis
Niloofar Arazkhani, Maciej Kotecki, Brent Cochran and Natasa Miskov-Zivanov 880

When Retrieval Doesn't Help: A Large-Scale Study of Biomedical RAG
Erfan Nourbakhsh, Rocky Slavin, Ke Yang and Anthony Rios 890

CrossDDI: Cross-Source Evidence-Grounded Drug-Drug Interaction Verification
Bohao Chu and Norbert Fuhr 911

GRAFT: Gated Retrieval-Augmented Fine-Tuning for Relation Extraction
Yuhang Jiang and Ramakanth Kavuluru 920

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 932

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 944

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 955

From Rules to Predictions: Federated Tabular Learning with LLM Reasoning
Afsaneh Mahanipour and Hana Khamfroush 970

MedBench: Deliberative Evaluation of Medical Language Models
Pratik Jalan, Mukul Joshi, Akhilesh Magotra and Kshitij Jadhav 981

Fast, Accurate, and Local Conversion of MIMIC-IV to OMOP with DBT
Adam Sutton, Niko Moller-Grell, Thomas Searle and Richard Dobson 992

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 997

FHexchange: Resources for Family Health History Extraction and Normalization From Consumer Dialog Sources
Michelle Nguyen, Nidhi Soley, Ayah Zirikly, João Sedoc and Casey Taylor 1014

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 1029

IndicMedDialog: A Parallel Multi-Turn Medical Dialogue Dataset for Accessible Healthcare in Indic Languages
Shubham Nigam, Suparnojit Sarkar and Piyush Patel 1041

Towards a Radiologist Imitation Framework for 3D CT Diagnosis with Multimodal LLMs
Kaidi Zhang, Zhiyuan Yan, Gao Cheng and Zhenyang Cai 1056

Probing and Steering Uncertainty in Biomedical Language Models: Representational Structure and Behavioral Limits
Debmalya Pal 1066

Relations of Linguistic Features and Medical Text Preferences are Nontrivial
Davis Bartels, Brandon Colelough and Dina Demner-Fushman 1080

Overview of the MedGenVidQA 2026 Shared Task on Medical Generative Video Question Answering
Deepak Gupta, Collin Campbell, Pedram Golnari and Dina Demner-Fushman 1089

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 1101

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*

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

Moustafa Yehia Hassan
Doha Institute for Graduate Studies
Doha, Qatar
moustafa.hassan@dohainstitute.edu.qa

Abstract

Computational mental health (CMH) classifiers often degrade under distribution shift because human annotators and distant-supervision pipelines reward different linguistic signals. We introduce TSS (Triple-Stream Stress probe), a multi-channel diagnostic framework that decomposes text into (A) lexical character n-grams, (B) a small, mostly content-free morpho-syntactic channel, and (C) a 154-feature psycholinguistic style channel. Across four English datasets ($N = 12{,}906$), TSS reveals a *lexical interference effect*: adding lexical features to the style channel reduces Macro-F1 on human-labeled data (mean drop 0.072, $p < 10^{-4}$) but not on auto-labeled data. We propose Degree of Divergence (DoD), a difference-in-differences statistic adapted from econometrics for label-source auditing, with instance-level bootstrap inference; the headline estimate is $\mathrm{DoD}_{\mathrm{BC\text{–}A}} = 0.0374$, 95% CI $[0.0097, 0.0651]$, $p = 0.0032$. A platform-stratified Twitter-only DoD (which removes the Reddit vs. Twitter contrast) reproduces the pattern with bootstrap inference: $\mathrm{DoD}^{\mathrm{Tw}}_{\mathrm{BC\text{–}A}} = +0.096$ ($p < 0.001$) and $\mathrm{DoD}^{\mathrm{Tw}}_{\mathrm{AC\text{–}A}} = -0.089$ ($p < 0.001$). Interventional masking (`pos_only`) retains ∼95–99% of Channel C's performance after destroying content words on human datasets, indicating that the style channel does not rely primarily on lexical surface form. TSS is positioned as a diagnostic audit framework, not a clinical screening tool: it flags label-source-specific shortcut learning before generalization claims are made.

1 Introduction

Computational Mental Health (CMH) promises scalable screening and monitoring from language, but its progress is constrained by a mismatch between benchmark labels and the construct they are taken to measure. Large corpora are frequently built via distant supervision—keyword filters, self-reported diagnoses, or community-membership proxies—which can systematically reward lexical cues (Coppersmith et al., 2015; Harrigian et al., 2021). Human annotators, in contrast, often rely on *how* something is said: fragmentation, negation scope, function-word usage, and stylistic signatures (Tausczik and Pennebaker, 2010; Pennebaker et al., 2003). When the two label sources are conflated, models can score well by exploiting spurious lexical correlations (*shortcut learning*; Geirhos et al., 2020) rather than learning robust psycholinguistic markers (Ernala et al., 2019; D'Amour et al., 2022).

Task. We study *binary stress detection* from short social-media posts. Each instance receives a label $y \in \{0, 1\}$, where $y{=}1$ indicates that the post expresses current psychological distress or stress-related experience as judged by the dataset's labeling protocol, and $y{=}0$ indicates the absence of such expression. Throughout this paper, "stress" refers to the dataset-level distress/stress label, *not* to a clinical diagnosis; no claim of diagnostic validity is made. The goal is not to build a state-of-the-art classifier, but to audit which linguistic signals different label sources reward.

Research questions. **RQ1 (Label-source divergence)** Does changing the label source (human vs. distant supervision) systematically change which linguistic channels are rewarded? **RQ2 (Lexical interference)** Does adding lexical content to stylistic features help or harm under human annotation? **RQ3 (Quantification)** Can the human-vs-auto gap be quantified as a single, statistically grounded scalar?

Contributions. (i) We release TSS, a channel-separable, length-robust auditing probe that decomposes text into lexical (A), mostly content-free morpho-syntactic (B), and psycholinguistic style (C) channels, with explicit length normalization and conditional scaling for cross-platform trans-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1–14
July 3-4, 2026 ©2026 Association for Computational Linguistics

fer. (ii) We document a *lexical interference effect*: on human-labeled data, augmenting C with lexical A degrades Macro-F1; an interventional masking suite (`pos_only`, `content_only`, `function_only`) tests whether performance survives deliberate destruction of lexical shortcuts. (iii) We introduce DoD, a difference-in-differences auditing statistic with instance-level bootstrap inference (n=10,000) and Benjamini–Hochberg FDR control, plus a *platform-stratified* variant that removes the Reddit vs. Twitter contrast. (iv) We provide leave-one-domain-out (LODO) evaluation, unsupervised stylistic phenotyping (bootstrap ARI ≈ 0.98), paired baseline comparisons against MentalBERT (Ji et al., 2022) and few-shot LLaMA-3 (Grattafiori et al., 2024; Brown et al., 2020), and a qualitative conflict-zone workbook. The full pipeline, decontamination scripts, and qualitative workbook are publicly available.[1]

2 Related Work

Psycholinguistic style. A long line of work shows that function words and stylistic patterns reflect cognitive style more reliably than consciously controlled topical content (Pennebaker et al., 2003; Tausczik and Pennebaker, 2010; Boyd et al., 2022). Cognitive therapy frameworks predict linguistically measurable correlates of distress, most notably absolutist language (Beck, 1976; Al-Mosaiwi and Johnstone, 2018) and specific cognitive distortions (Shickel et al., 2020).

Distant supervision and shortcut learning in CMH. Distant-supervision pipelines built from self-reported diagnoses (Coppersmith et al., 2015) or community membership have repeatedly been shown to yield models that rely on surface lexical cues and degrade under distribution shift (Harrigian et al., 2021; Aguirre et al., 2021; Ernala et al., 2019). These failures align with the broader finding that neural classifiers tend to latch onto annotation artifacts and spurious correlations (Geirhos et al., 2020; Gururangan et al., 2018; McCoy et al., 2019; D'Amour et al., 2022), and that distributional robustness requires explicit intervention (Sagawa et al., 2020; Kaushik et al., 2020). Our work differs in goal: rather than proposing a more robust classifier, we propose a *diagnostic framework* that quantifies how much a given dataset's labels reward lexical shortcuts, in the spirit of behavioral auditing (Ribeiro et al., 2020).

Stress detection benchmarks. Our experimental backbone uses Dreaddit (Turcan and McKeown, 2019) as the structured human-labeled benchmark and MentalBERT (Ji et al., 2022) as a representative domain-pretrained transformer baseline. We also draw on the Twitter mental-health corpora derived from Coppersmith et al. (2015), treated here as a distant-supervision proxy rather than a stress gold standard (see §3).

3 Data and Task

3.1 Datasets

We evaluate on four English datasets after within-dataset exact-match deduplication. Table 1 summarizes the platform, size, label source, and intended use of each dataset.

3.2 Twitter-auto as a Proxy, Not Gold

Twitter-auto is not used as a gold stress dataset. It is constructed by treating posts whose authors self-reported a mental-health diagnosis (depression, PTSD, anxiety, following Coppersmith et al., 2015) as *auto-positive*, under the working assumption that clinically diagnosed populations exhibit elevated linguistic distress markers. We acknowledge that this introduces a *disorder→stress conflation*: a diagnosed user is not necessarily expressing distress in every post, and the resulting label is a noisy proxy for the post-level stress construct studied here. We treat Twitter-auto purely as a distant-supervision proxy for testing whether such labels reward different linguistic signals than human distress annotations. Twitter-gold is a manually annotated subsample drawn from the same underlying corpus, where annotation targeted *perceived psychological distress* rather than the presence of mental-health keywords; stress-keyword removal was applied to reduce trivial lexical leakage. We do not claim expert-clinical annotation for this set; methodological constraints of the manual-validation protocol are acknowledged in §7. A related, substantially smaller manually annotated Twitter validation subset (120 examples) was constructed by Rastogi et al. (2022) to validate automated annotation strategies, which we cite here as methodological context rather than as the same artifact.

[1] Code and appendix: `https://github.com/MoustafaMohamedMoustafaHassan/TSS-Probe-CMH`

Dataset	Platform	N	Label source	Positive label means	Used for
Dreaddit-test	Reddit	715	Human (expert)	Stress / distress expression	Human-labeled evaluation
Twitter-gold	Twitter	2,863	Human (manual annotation)	Perceived psychological distress	Human-labeled evaluation
Twitter-auto	Twitter	6,218	Distant supervision	Mental-health / distress *proxy*	Auto-label comparison
Reddit-combi	Reddit	3,110	Distant supervision	Community-/keyword-derived proxy	Auto-label comparison

Table 1: Dataset overview. Positive prevalence: Dreaddit-test 0.516, Twitter-gold 0.338, Twitter-auto 0.478, Reddit-combi 0.880 (i.e., the positive label is the *majority* class on Reddit-combi; minority prevalence = 0.120). Dreaddit follows Turcan and McKeown (2019); Twitter-auto and the Twitter-gold subsample derive from Coppersmith et al. (2015); Reddit-combi is a combined Reddit corpus assembled under distant supervision (see §3.3).

3.3 Reddit-combi as a Distant-Supervision Proxy

Reddit-combi is an internally assembled Reddit corpus constructed under distant supervision and released with the artifact bundle as `data/raw/Reddit_Combi.csv` (with title/body fields and binary labels). It is not an externally published benchmark; it is an artifact-level distant-supervision corpus assembled for this study and released with the repository, and we report its provenance through the artifact path. Because its construction is not equivalent to expert annotation, we use Reddit-combi exclusively as a Reddit-side auto-label proxy and pair it with Twitter-auto for the auto-source half of the DoD contrast. Crucially, the positive label is the *majority* class on Reddit-combi (0.880); we therefore favor Macro-F1, PR-AUC, and balanced accuracy over raw accuracy or class-1 F_1 when reporting this dataset, and we never use Reddit-combi as a stand-in for human annotation.

3.4 Decontamination and Imbalance

To prevent leakage in cross-platform CMH evaluation, we perform exact-match cross-dataset decontamination on `cleaned_text`, removing overlapping posts from the training pool while preserving evaluation sets. The datasets exhibit severe class imbalance, especially Reddit-combi, where the positive label is the *majority* class (positive prevalence 0.880; minority prevalence 0.120). Accordingly we report Macro-F1 as the primary metric, complemented by PR-AUC under skewed prevalence (PR-AUC is preferred over ROC-AUC when one class dominates). A naive majority-class baseline yields Macro-F1 of 0.340, 0.343, 0.398, and 0.468 on Dreaddit-test, Twitter-auto, Twitter-gold, and Reddit-combi respectively; all TSS channels substantially exceed these baselines, confirming signal acquisition beyond prevalence effects. For paired comparisons we use McNemar's test (McNemar, 1947) on discordant predictions and paired bootstrap CIs (Efron, 1979) for Macro-F1 differences, controlling false discoveries with Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995).

Dreaddit additionally provides five heterogeneous stressor domains (abuse, anxiety, financial, PTSD, social), enabling leave-one-domain-out (LODO) evaluation that probes cross-stressor generalization beyond platform transfer.

4 The TSS Probe

TSS is a diagnostic decomposition, not a competitive classifier. The three channels are designed so that their contrast is informative: comparing what each rewards under each label source isolates lexical-shortcut behavior.

4.1 Channel A: Lexical Surface Form

Channel A uses character-level TF–IDF n-grams of length 3–5 (Salton and Buckley, 1988), with χ^2 feature selection at $k = 500$ (Yang and Pedersen, 1997). We use character n-grams over word n-grams for two reasons. First, social-media text contains hashtags, contractions, emoji, intentional misspellings, and morphological variation that destabilize word tokenization; character-level features are robust to these (Coppersmith et al., 2015). Second, we want Channel A to provide a strong, sub-word lexical baseline so that any *lexical interference effect* we observe cannot be trivially attributed to vocabulary coverage gaps in a word-based representation. The χ^2 filter at k=500 prevents dimensionality blow-up and is comparable in size to standard sparse lexical baselines for short-text classification.

4.2 Channel B: Mostly Content-Free Morpho-Syntax

Channel B converts POS bigrams and abstract POS-SVO triples (extracted over the first 500 characters per document for efficiency, using the Penn Treebank tagset via spaCy; Honnibal et al., 2020) into six interpretable features: five POS-derived plus one absolutist-ratio (B_{abs}), which is the only Channel B feature that accesses lexical tokens (a closed absolutist-word list, following Al-Mosaiwi and Johnstone, 2018). Channel B uses adaptive length smoothing and log-dampened mass features for cross-platform robustness; the exact formulations are given in Appendix A.

Channel B is a diagnostic, not a performance booster. We do *not* present Channel B as an ingredient that improves accuracy when combined with C. As shown in §5, adding B to C is approximately neutral on most datasets (occasionally negative on Dreaddit-test). This is itself informative: C already absorbs much of the useful structural variation; B is useful as (i) a near-content-free, privacy-oriented diagnostic (5 of 6 features discard lexical tokens), and (ii) a structural stress test that probes whether performance survives a near-complete loss of vocabulary.

4.3 Channel C: Psycholinguistic Style

Channel C is the principal stylistic representation, with 154 features grouped into seven feature families (Table 2). It applies a length-robust squashing transform $c_{\text{out}} = \tanh(2c)$ to convert continuous features to soft binary triggers resistant to cross-platform distributional shift, and uses Yule's I rather than TTR for lexical diversity because TTR mathematically decays with text length (Appendix A).

Beyond sentiment-like lexicons, Channel C explicitly encodes cognitive fragmentation (a hallmark of anxiety) via the coefficient of variation of sentence length, short-burst ratios, and abrupt punctuation patterns, as well as structural negation scope distinguishing helplessness (NEG→VERB) from negative self-evaluation (NEG→ADJ).

4.4 Unified Classifier and Calibration

All channels are trained with linear models for interpretability. Feature vectors are L_2-normalized: $x' = x/\|x\|_2$. Conditional scaling disables `StandardScaler` for any channel containing B, to avoid reintroducing length bias. Class imbalance is handled by `class_weight='balanced'` (Pedregosa et al., 2011) rather than undersampling. Decision thresholds are not fixed at 0.5: we calibrate by maximizing F_1 over a bounded grid $[0.20, 0.80]$ on out-of-fold (OOF) probabilities (no leakage). Dense lexical/style channels use L_2 (ridge); any channel containing B uses Elastic-Net (Zou and Hastie, 2005) with internal 5-fold CV; the selected ℓ_1-ratio converges to 0.95, indicating an algorithmic preference for sparse structural evidence.

Feature block	Examples	#
Base style / readability / POS-ratio	Flesch–Kincaid, sentence-length CV, function-word ratios, pronoun densities	29
Open lexicon categories	affect, cognitive, social process, sensory; LIWC-style counts and ratios	110
Structural rhythm / negation / punctuation	ellipses, "!?", NEG→VERB, NEG→ADJ, scope balance, short-burst ratio	11
Raw intensity + Yule's I	caps ratio, elongated tokens, punctuation density, length-independent diversity	4
Total		**154**

Table 2: Channel C feature blocks with feature counts as implemented in `ChannelC_Extended` of the released codebase. Open lexicon categories follow LIWC-style groupings (Tausczik and Pennebaker, 2010; Boyd et al., 2022); absolutist markers follow Al-Mosaiwi and Johnstone (2018); negation-scope features operationalize the helplessness vs. self-evaluation distinction motivated by Beck (1976).

4.5 Degree of Divergence (DoD)

DoD is a difference-in-differences statistic adapted from causal econometrics to quantify label-source bias in NLP (Appendix A, Eq. 3):

$$\text{DoD} = \Delta_h - \Delta_a \quad (1)$$

where each Δ compares a structural channel (e.g., BC) against lexical A within the same label source. DoD is estimated with an instance-level bootstrap ($n = 10{,}000$) over all 12,906 instances. Because DoD as introduced above mixes label source with platform (human Reddit/Twitter vs. auto Reddit/Twitter), we additionally compute a *platform-stratified Twitter-only* DoD that removes the Reddit vs. Twitter contrast (§5.4).

4.6 Masking Suite

Five masks are evaluated: `none`, `pos_only`, `content_only`, `function_only`, `random_pos` (negative control). Unlike standard ablations that only remove feature blocks, the masking suite intervenes directly on the text: `pos_only` destroys lexical content while preserving syntactic traces; `content_only` preserves topical content while erasing function words and syntactic cues; `function_only` isolates function-word dynamics; `random_pos` acts as a destructive control. This design probes whether a channel's performance survives deliberate destruction of lexical shortcuts, complementing the observational DoD analysis.

5 Results

We first summarize per-channel Macro-F1 across human and auto label sources, then quantify lexical interference, DoD, and a platform-stratified Twitter-only DoD; finally we report interventional masking, baselines, and robustness audits.

5.1 Channel Performance

Table 3 and Table 4 report Macro-F1 per channel on human-labeled and auto-labeled datasets respectively. Two patterns stand out. First, the structural / style combination BC achieves nearly identical Macro-F1 on the two human-labeled datasets (0.690 on Dreaddit-test vs. 0.690 on Twitter-gold; $\Delta \approx 0.0002$), suggesting low degradation under platform shift when relying on structure and style. Second, moving from human-labeled Twitter-gold to auto-labeled Twitter induces a sharp performance drop in C and BC (e.g., C: $0.701 \rightarrow 0.504$; BC: $0.690 \rightarrow 0.455$). From an auditing standpoint, this drop is consistent with *label-source divergence*: auto labels reward lexical proxy cues that structural and style probes are less willing to mimic.

5.2 Lexical Interference Effect

Table 5 reports the instance-bootstrap test for the lexical interference effect, defined as the Macro-F1 change when augmenting C with A. On human-labeled data, the mean drop is 0.072 (95% CI $[0.052, 0.092]$, one-sided $p < 10^{-4}$); on auto-labeled data, the change is -0.020 (95% CI $[-0.033, -0.007]$, one-sided $p = 0.9988$ for "drop > 0"). All instance-level tests are paired (same instances) and corrected via Benjamini–Hochberg FDR (27 tests, 14 rejections).

Channel	Dreaddit (H) [CI]	Tw-G (H) [CI]
A	0.634 [.598,.669]	0.661 [.643,.678]
AB	0.632 [.596,.667]	0.650 [.631,.669]
B	0.561 [.525,.598]	0.586 [.566,.605]
C	**0.739 [.706,.771]**	**0.701 [.682,.719]**
AC	0.679 [.644,.713]	0.617 [.599,.635]
BC	0.690 [.655,.724]	0.690 [.671,.709]
ABC	0.690 [.655,.724]	0.689 [.670,.708]

Table 3: Macro-F1 by channel on human-labeled datasets with 95% bootstrap CIs (Channel C in bold). Tw-G = Twitter-gold; leading-decimal abbreviation in CIs.

Channel	Tw-A (A) [CI]	Red-c (A) [CI]
A	0.522 [.510,.535]	0.611 [.587,.635]
AB	0.460 [.448,.472]	0.589 [.564,.613]
B	0.463 [.451,.475]	0.579 [.554,.604]
C	0.504 [.492,.516]	0.681 [.658,.704]
AC	0.567 [.555,.580]	0.658 [.634,.681]
BC	0.455 [.444,.467]	0.689 [.665,.712]
ABC	0.455 [.444,.467]	0.689 [.665,.712]

Table 4: Macro-F1 by channel on auto-labeled datasets with 95% bootstrap CIs. Tw-A = Twitter-auto, Red-c = Reddit-combi; leading-decimal abbreviation in CIs.

5.3 Degree of Divergence

Table 6 reports the full DoD estimates. $\mathrm{DoD}_{\mathrm{C–A}} = +0.047$ and $\mathrm{DoD}_{\mathrm{AC–A}} = -0.045$ form a symmetric pair: human annotators implicitly weight style, while auto labels reward lexical proxies. The negative $\mathrm{DoD}_{\mathrm{AC–A}}$ is the strongest quantitative signature of the lexical interference effect: lexical augmentation actively penalizes human-source performance.

A note on the permutation diagnostic. A regime-permutation diagnostic (Good, 2005) gives $p = 0.092$ (FDR-adjusted $p = 0.1656$). We treat this as a conservative robustness audit: the bootstrap evaluates stability across instances, whereas the permutation diagnostic probes whether the observed channel ranking could arise under channel-design constraints. We prioritize the bootstrap CI for effect estimation while reporting the permutation result transparently.

5.4 Platform-Stratified Twitter-Only DoD

A potential confound is that human-labeled and auto-labeled datasets differ not only in label source but also in platform composition. The standard DoD above assumes that platform effects are additive and channel-invariant within each label source. We test this directly by computing a *Twitter-only*

Source (C → AC)	Mean drop	95% CI	p (1-s.)
Human	0.0720	$[0.052, 0.092]$	$< 10^{-4}$
Auto	-0.0201	$[-0.033, -0.007]$	0.9988

Table 5: Lexical interference test (instance bootstrap). Augmenting Channel C with Channel A reduces Macro-F1 on human labels and slightly *increases* it on auto labels.

DoD	Est.	95% CI	p	d
BC–A	$+0.0374$	$[0.010, 0.065]$	0.0032	2.63
C–A	$+0.0469$	$[0.020, 0.075]$	0.0001	3.32
AC–A	-0.0453	$[-0.068, -0.022]$	<0.001	-3.89
AB–A	$+0.0360$	$[0.010, 0.062]$	0.0036	2.68

Table 6: DoD estimates (instance bootstrap, n=10,000). For positive DoD the alternative is DoD > 0; for $\mathrm{DoD}_{\text{AC–A}}$ the alternative is DoD < 0.

DoD that holds the platform fixed (Twitter-gold vs. Twitter-auto) and removes the Reddit vs. Twitter contrast:

$$\mathrm{DoD}^{\text{Tw}}_{X-A} = [\, F_1^{\text{Tw-gold}}(X) - F_1^{\text{Tw-gold}}(A) \,] - [\, F_1^{\text{Tw-auto}}(X) - F_1^{\text{Tw-auto}}(A) \,]. \quad (2)$$

We attach uncertainty to this platform-stratified check via a paired bootstrap that resamples each Twitter dataset independently with replacement (N_{boot}=2,000; Table 7). The effect remains stable and highly significant: $\mathrm{DoD}^{\text{Tw}}_{\text{C–A}} = +0.058$ (95% CI $[+0.031, +0.086]$, $p < 0.001$), $\mathrm{DoD}^{\text{Tw}}_{\text{BC–A}} = +0.096$ ($[+0.066, +0.125]$, $p < 0.001$), and $\mathrm{DoD}^{\text{Tw}}_{\text{AC–A}} = -0.089$ ($[-0.113, -0.065]$, $p < 0.001$). The signs and magnitudes match the cross-platform DoD: structural channels gain relative to A under human labels and lose under auto labels, even when both label sources come from the same platform. This does not eliminate every label-construction difference (e.g., annotation protocol, sampling strategy), but it makes a pure platform-shift explanation unlikely. A controlled within-platform dual-annotation experiment would still be required for definitive causal identification (§7).

5.5 Interventional Masking

Table 8 reports masking-suite Macro-F1 on human-labeled datasets, including 95% bootstrap CIs and paired bootstrap CIs for the difference vs. `none` (N_{boot}=2,000). On Dreaddit-test, `C/pos_only` is statistically indistinguishable from `C/none` ($\Delta = -0.006$, 95% CI $[-0.037, +0.023]$, $p = 0.72$), supporting the interpretation that Channel C's signal is not primarily carried by lexical surface form. On Twitter-gold, the same condition shows a small but reliable drop ($\Delta = -0.034$, 95% CI $[-0.052, -0.017]$, $p < 0.001$), consistent with noisier, shorter texts where POS traces carry less recoverable information after lexical deletion. Notably, `BC/pos_only` on Dreaddit *exceeds* unmasked `BC/none` ($\Delta = +0.037$, 95% CI $[+0.008, +0.066]$, $p = 0.013$): destroying content under the combined channel forces the classifier off syntactically irrelevant morpho-lexical patterns and onto genuinely structural ones. `random_pos` acts as a destructive control; on Twitter-gold it produces large significant drops for both C ($\Delta = -0.080$, $p < 0.001$) and BC ($\Delta = -0.062$, $p < 0.001$), confirming that performance does not survive arbitrary syntactic scrambling.

Contrast	DoD^{Tw} [95% CI]	p
C – A	$+0.058$ $[+0.031, +0.086]$	<0.001
BC – A	$+0.096$ $[+0.066, +0.125]$	<0.001
AC – A	-0.089 $[-0.113, -0.065]$	<0.001
AB – A	$+0.051$ $[+0.022, +0.080]$	<0.001

Table 7: Platform-stratified Twitter-only DoD with 95% bootstrap CIs and two-sided p-values (N_{boot}=2,000). DoD^{Tw} is the difference of within-source $\Delta = F_1(\text{channel}) - F_1(A)$ between Twitter-gold (human) and Twitter (auto). Signs and magnitudes match the cross-platform DoD, making a pure platform-shift explanation unlikely.

Channel / mask	Dreaddit F1 [CI]	Tw-G F1 [CI]
C / none	0.739 [.708,.772]	0.701 [.682,.719]
C / pos_only	0.734 [.700,.767]	0.666 [.647,.685]
C / function_only	0.736 [.705,.769]	0.603 [.582,.624]
C / content_only	0.724 [.691,.756]	0.652 [.632,.671]
C / random_pos	0.742 [.709,.774]	0.621 [.599,.641]
BC / none	0.690 [.654,.724]	0.690 [.670,.708]
BC / pos_only	0.728 [.694,.760]	0.594 [.574,.614]
BC / function_only	0.727 [.692,.761]	0.665 [.645,.683]
BC / content_only	0.724 [.692,.754]	0.625 [.606,.644]
BC / random_pos	0.739 [.706,.771]	0.628 [.607,.649]

Table 8: Masking-suite Macro-F1 on human-labeled datasets with 95% bootstrap CIs (cell-level, N_{boot}=2,000). Tw-G = Twitter-gold; CIs use leading-decimal abbreviation (e.g., [.708,.772] ≡ [0.708, 0.772]). Paired CIs for Δ vs. `none` are in Table 9.

5.6 Baselines as Diagnostics

Table 10 reports paired comparisons against MentalBERT (Ji et al., 2022) and 3-shot LLaMA-3-8B (Grattafiori et al., 2024; Brown et al., 2020). LLaMA-3 is evaluated only in a few-shot diag-

Comparison	Dreaddit Δ [95% CI]	p	Tw-gold Δ [95% CI]	p
C: pos_only − none	-0.006 [$-0.037, +0.023$]	0.72	-0.034 [$-0.052, -0.017$]	<0.001
C: function_only − none	-0.003 [$-0.031, +0.025$]	0.88	-0.098 [$-0.118, -0.078$]	<0.001
C: content_only − none	-0.016 [$-0.041, +0.010$]	0.20	-0.049 [$-0.063, -0.033$]	<0.001
C: random_pos − none	$+0.003$ [$-0.018, +0.025$]	0.77	-0.080 [$-0.096, -0.063$]	<0.001
BC: pos_only − none	$+0.037$ [$+0.008, +0.066$]	0.013	-0.096 [$-0.112, -0.081$]	<0.001
BC: function_only − none	$+0.037$ [$+0.002, +0.071$]	0.039	-0.025 [$-0.045, -0.007$]	0.010
BC: content_only − none	$+0.033$ [$+0.005, +0.062$]	0.020	-0.064 [$-0.080, -0.049$]	<0.001
BC: random_pos − none	$+0.049$ [$+0.023, +0.076$]	0.001	-0.062 [$-0.077, -0.047$]	<0.001

Table 9: Paired bootstrap differences for masking conditions vs. `none` (N_{boot}=2,000). On Dreaddit, Channel C is statistically unaffected by lexical destruction; on Twitter-gold, the drops are significant but small in absolute terms (95–96% retention for `pos_only`). The BC pos_only case on Dreaddit shows that combined structure-plus-style representation can *benefit* from lexical destruction.

Dataset	Model	Macro-F1
Dreaddit (H)	TSS-C	0.739
Dreaddit (H)	TSS-BC	0.690
Dreaddit (H)	MentalBERT	0.801
Dreaddit (H)	LLaMA-3 (3s)	0.613
Tw-gold (H)	TSS-C	0.701
Tw-gold (H)	TSS-BC	0.690
Tw-gold (H)	MentalBERT	0.715
Tw-gold (H)	LLaMA-3 (3s)	0.800
Tw-auto (A)	TSS-C	0.504
Tw-auto (A)	TSS-BC	0.455
Tw-auto (A)	MentalBERT	0.418
Tw-auto (A)	LLaMA-3 (3s)	0.522
Reddit-c. (A)	TSS-C	0.681
Reddit-c. (A)	TSS-BC	0.689
Reddit-c. (A)	MentalBERT	0.760
Reddit-c. (A)	LLaMA-3 (3s)	0.794

Table 10: Baseline comparison (Macro-F1). Full paired ΔF_1, bootstrap CIs, and McNemar p-values are in Appendix B.3.

nostic capacity. Its instability is informative (Appendix C.5): on Dreaddit-test recall is 0.984 with precision 0.606 (over-firing); on Twitter-auto recall collapses to 0.203 (under-firing). We interpret this as evidence that few-shot lexical priors track the availability of stress vocabulary rather than the underlying state.

We treat baselines as epistemic diagnostics rather than direct competitors: when MentalBERT outperforms TSS on human labels, it plausibly leverages legitimate contextual signal; when both strong baselines excel on auto-labeled corpora, they may be matching label-source-induced proxy noise; TSS's relative stability across label sources is the central observation. On Twitter-auto, C and BC outperform MentalBERT yet remain below LLaMA-3; on Twitter-gold MentalBERT is approached by TSS-C ($\Delta Macro - F1 = -0.0141$, paired bootstrap 95% CI $[-0.0367, +0.0082]$, $p = 0.108$; Appendix B.3), but LLaMA-3-8B (3-shot) attains a higher Macro-F1 (0.800). We therefore *do not* claim overall competitive superiority on Twitter-gold; we read the result as evidence of representation efficiency relative to one domain-pretrained transformer baseline, not as a SOTA claim.

5.7 LODO and Phenotyping

Under leave-one-domain-out evaluation on Dreaddit's five stressor domains (Appendix C.1), Channel C has the highest mean out-of-domain Macro-F1 among TSS channels (0.685) with notably low variance (std 0.015), supporting a relative cross-stressor invariance claim. MentalBERT attains a higher mean (0.818) but with greater variance (std 0.030), consistent with domain-dependent lexical sensitivity.

K-Means (Lloyd, 1982; MacQueen, 1967) with K=3 on Channel C yields highly stable clusters (bootstrap ARI 0.980, 95% CI $[0.961, 0.991]$, n_{boot}=100). Cluster–dataset association (AMI 0.281, p=0.0005) flags partial platform entanglement, an explicit leakage audit reinforcing the need for label-source-aware evaluation. Channel orthogonality is measured rather than assumed (mean $|\rho| \approx 0.595$); the weakest pair (A vs. B, $|\rho| = 0.296$) supports the intended factorization. Channel C trains in ≈ 27.5 s on a single CPU (Appendix C.4), supporting the diagnostic-density framing.

6 Discussion

TSS is best understood as a *diagnostic tool*: it quantifies when results are label-source-specific and provides a single statistically grounded scalar (DoD) to summarize that gap. Below we consolidate the interpretation, formalize when "strong baselines"

are mostly a lexical illusion, and position structural invariants as the signal that survives label-source shifts.

6.1 Why Lexical Baselines Look Strong

In auto-labeled or closed-community settings, topical stress lexicon co-occurs with structural distress markers, producing a *semantic–structural intersection (SSI)*. A model can succeed by detecting topics that correlate with distress rather than the distress state itself, consistent with shortcut learning under non-causal correlation (Geirhos et al., 2020). This reconciles the high performance of MentalBERT (Macro-F1 0.760) and LLaMA-3 (Macro-F1 0.794) on Reddit-combi with their pronounced degradation on Twitter-auto (MentalBERT Macro-F1 0.418), where SSI is weaker and intent is mixed (news, ads, meta-talk; see qualitative cases in Appendix D).

A direct signal-saturation test on Reddit-combi: augmenting C with morpho-syntax yields no detectable gain ($+0.008$ Macro-F1; $p = 0.259$; Appendix B.1), consistent with SSI-induced overlap. The fact that C still beats A on Reddit-combi by $+0.070$ ($p < 0.0001$; Appendix B.4) is best read as an SSI-saturation effect: under high SSI, all channels benefit; the diagnostic question is the *relative* channel advantage *across* label sources, which is exactly what DoD captures and confirms ($\mathrm{DoD}_{\mathrm{C–A}} = +0.047$, $p < 10^{-4}$).

6.2 Representation Efficiency, Not SOTA

On Twitter-gold, the gap between TSS-C and MentalBERT is small ($\Delta Macro - F1 = -0.0141$, 95% CI $[-0.0367, +0.0082]$, $p = 0.108$; Appendix B.3)—consistent with, but not proving, statistical equivalence relative to *this* baseline. We do *not* claim that TSS-C matches the strongest baseline overall: LLaMA-3-8B (3-shot) reaches Macro-F1 0.800. The reading is *representation efficiency*: a 154-feature, fully interpretable linear channel approaches one domain-pretrained transformer on human-labeled data with radically lower representational and computational cost.

6.3 Implicit Stress vs. Keyword-Driven False Positives

Conflict-zone examples (Appendix D) instantiate two complementary failure modes: *keyword-driven false positives* (stress meta-talk, psychoeducational content, or promotional posts classified as distress because stress words are present even when structural distress is absent), and *implicit stress* (distress expressed without explicit stress lexicon, via hedging, negation, or fragmentation). At the instance level, across 800 audited cases Channel C alone corrects 253 Channel-A errors (195 keyword-driven false positives + 58 implicit-stress false negatives) and BC corrects 241, showing that lexical shortcuts fail in two directions and that structural style features recover both. TSS audits state inference, not topic detection.

6.4 Cross-Stressor Invariance and Clinical Bridge

LODO results (Appendix C.1) show Channel C varies little across held-out stressor domains (std 0.015), supporting the view that structural and stylistic markers track a cognitive–affective state that persists across heterogeneous causes; cognitive fragmentation, absolutist appraisal, and negation scope remain candidate distress signatures (Beck, 1976; Al-Mosaiwi and Johnstone, 2018). We use social-media data only as a stress-test environment for label-source bias, not as a clinical stand-in; analogous shortcut-learning risks arise in clinical NLP (Ernala et al., 2019), and TSS is compatible with privacy-constrained workflows because Channel B is mostly de-lexicalized.

7 Conclusion

We introduced TSS, a diagnostic framework for testing whether different label sources reward different linguistic evidence in computational mental health classification. Across four English datasets, the central result is a *lexical interference effect*: adding lexical surface features to style features reduces Macro-F1 on human-labeled datasets (mean drop 0.072, $p < 10^{-4}$) but not on auto-labeled datasets, with the gap quantified by $\mathrm{DoD}_{\mathrm{BC–A}} = 0.0374$ ($p = 0.0032$). Platform-stratified Twitter-only DoD (Table 7, all contrasts $p < 0.001$) and interventional masking on Dreaddit, where Channel C is statistically unaffected by full deletion of content words ($\Delta = -0.006$, 95% CI $[-0.037, +0.023]$, $p = 0.72$), suggest the effect is not reducible to ordinary platform shift; a controlled within-platform dual-annotation study remains necessary for causal identification. TSS is therefore an audit workflow that flags label-source-specific shortcut learning before generalization claims are made.

Limitations

Data and annotation. Evaluation is restricted to four English datasets and to a binary stress label. For Twitter-gold, inter-annotator agreement (Cohen's κ) was not computed under the manual-annotation protocol available to us; this limits claims about "human annotation" as a general construct. Twitter-auto inherits a disorder→stress conflation from Coppersmith et al. (2015), which is itself a form of distant-supervision noise.

Confounding between label source and platform. The four datasets cross two label sources with two platforms but do not exhaust the design. The platform-stratified Twitter-only DoD (§5.4) removes the most obvious platform confound, but a fully controlled within-platform dual-annotation experiment is required for causal identification of label-source effects.

Feature-level ablation and masking control. Per-cell bootstrap CIs for the masking suite are now reported (Tables 8–9). We do not perform a feature-level within-channel ablation: TSS is presented as a family-level diagnostic decomposition, and a full feature-attribution study is left for future work. We treat feature-family ablation as a separate attribution study rather than a camera-ready addition, because the central claim concerns channel-level label-source divergence rather than ranking individual features within Channel C. This does not affect the channel-level claim tested here, but it limits feature-level interpretability inside Channel C. The `random_pos` negative control does not collapse to chance because POS distributional cues remain after shuffling.

Word- vs. character-level lexical baseline. We did not include a word-unigram/bigram lexical baseline in this camera-ready version; Channel A is intentionally character-based to stress-test lexical surface dependence under noisy social-media spelling. Future work should compare word- and character-level lexical baselines directly to verify that the *lexical interference effect* is not specific to sub-word tokenization.

Clinical translation. Social-media data are not clinical data. We make no claim that TSS is a diagnostic medical tool; any deployment in a clinical setting requires clinician oversight, informed consent, and risk management. DoD may also require calibration for morphologically richer languages (e.g., Arabic, Chinese), where surface-form variation interacts differently with character-level lexical representations.

Few-shot baseline. LLaMA-3-8B is reported only in a 3-shot diagnostic capacity; instability is interpreted as evidence about few-shot lexical priors, not as a comment on fine-tuned LLM performance.

Ethical Considerations

This work is not a diagnostic medical tool. It focuses on auditing label bias and improving methodological rigor in CMH NLP. All evaluations use previously collected research datasets or released artifact-level data; no new user data was collected for this study. We use qualitative examples only after redacting personally identifying details. Any deployment of distress-detection systems built on social-media text requires clinician oversight, informed consent, and careful risk management, particularly given documented disparities in CMH model behavior across demographic groups (Aguirre et al., 2021).

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2932–2949.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

Aaron T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and Prafulla Dhariwal. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10. Association for Computational Linguistics.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, and Alex Beutel. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, and Matthias Bethge. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Phillip I. Good. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edition. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. `https://doi.org/10.5281/zenodo.1212303`.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, and Olivier Grisel. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Aryan Rastogi, Qian Liu, and Erik Cambria. 2022. Stress detection from social media articles: New dataset benchmark and analytical study. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107. Association for Computational Linguistics.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Appendices

Sections A–F below provide mathematical formulations, full statistical tables, robustness audits, qualitative case analysis, extended metrics, and SHAP audits referenced in the main text.

A Mathematical Formulations

A.1 Degree of Divergence

$$\begin{aligned}\mathrm{DoD} \;=\; &\big[M_h(BC) - M_h(A)\big] \\ &-\; \big[M_a(BC) - M_a(A)\big]\end{aligned} \tag{3}$$

DoD is a difference-in-differences statistic adapted from causal econometrics; M_h and M_a denote Macro-F1 under human and auto label sources respectively. Inference is by instance-level bootstrap (n=10,000).

A.2 Adaptive Length Smoothing (Channel B)

$$\lambda \;=\; \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\,\min(L/L_{\text{ref}},\,1) \tag{4}$$

with $\lambda_{\min}$=3.0, $\lambda_{\max}$=20.0, L_{ref}=50.0. Short texts retain sparse POS signals; long texts are regularized.

A.3 Length-Normalized Log-Odds Mass (Channel B)

$$B_{\text{pos_mass}} \;=\; \frac{\log(1+M_{\text{pos}})}{\log(1+L_{\text{eff}})}, \qquad L_{\text{eff}} = L + \lambda \tag{5}$$

The remaining Channel B features are $B_{\text{polarity}} = M_{\text{pos}}/(M_{\text{pos}}+M_{\text{neg}}+\varepsilon)$, $B_{\text{load}} = \log_e(1+M_{\text{pos}}+M_{\text{neg}})/\max(\log_e(1+\text{cnt}),\,1)$, and $B_{\text{abs}} = |\{t : t \in \mathcal{A}\}|/L_{\text{eff}}$, where $\mathcal{A}$ is a closed absolutist-word list (Al-Mosaiwi and Johnstone, 2018). Of the six Channel B features, B_{abs} is the only one that accesses lexical tokens.

A.4 Structural Negation Balance (Channel C)

$$\text{NegBalance} \;=\; \frac{R_{\text{Neg}\to\text{Verb}} - R_{\text{Neg}\to\text{Adj}}}{N_{\text{neg_verb}} + N_{\text{neg_adj}} + \varepsilon} \tag{6}$$

Operationalizes the helplessness vs. negative self-evaluation distinction motivated by Beck (1976).

A.5 Length-Robust Squashing (Channel C)

$$c_{\text{out}} \;=\; \tanh(2c) \tag{7}$$

A.6 Length-Independent Lexical Diversity

$$I \;=\; \frac{{M_1}^2}{M_2 - M_1} \tag{8}$$

where M_1 is the vocabulary size and $M_2 = \sum_i f_i^2$ is the sum of squared per-type frequencies. This formulation follows the convention used in our pipeline (`features.py`, `_calc_yule_i`); values are capped at 10^3 to prevent extreme outliers from very short texts. Yule's I is preferred over TTR because TTR decays mathematically with text length, biasing cross-platform comparison from long Reddit posts to short tweets.

B Detailed Statistics

B.1 SSI Saturation Test (Reddit-combi)

Setting	ΔF1 (BC−C)	p
Reddit-combi (Auto)	+0.0080	0.259

Table 11: SSI signal-saturation equivalence test (paired instance bootstrap).

B.2 Representation Efficiency Note

The single-row paired-bootstrap result for TSS-C vs. MentalBERT on Twitter-gold ($\Delta Macro-F1 = -0.0141$, 95% CI $[-0.037, +0.008]$, $p = 0.108$) appears in the top row of Table 12. The CI does not exclude zero, but LLaMA-3 (3-shot)

reaches Macro-F1 0.800 on the same dataset, so this is *not* a SOTA claim.

B.3 Paired Baseline Comparisons

Pair (dataset)	ΔF1	95% CI
C / MB (Tw-G)	-0.014	$[-0.037, +0.008]$
C / LLaMA (Tw-G)	-0.099	$[-0.121, -0.078]$
C / MB (Tw-A)	$+0.086$	$[+0.073, +0.099]$
C / LLaMA (Tw-A)	-0.018	$[-0.032, -0.004]$
BC / MB (Tw-G)	-0.025	$[-0.045, -0.004]$
BC / LLaMA (Tw-G)	-0.110	$[-0.132, -0.088]$
BC / MB (Tw-A)	$+0.037$	$[+0.027, +0.048]$
BC / LLaMA (Tw-A)	-0.067	$[-0.080, -0.054]$

Table 12: Paired baseline comparisons (paired bootstrap CI). Pairs are TSS channel vs. baseline. MB = MentalBERT; LLaMA = LLaMA-3-8B (3-shot); Tw-G = Twitter-gold; Tw-A = Twitter-auto.

Pair	p_{McN}	d_z
TSS-C vs. MB (Tw-gold)	4.49×10^{-5}	-1.47
TSS-C vs. LLaMA-3 (Tw-gold)	2.55×10^{-15}	-11.40
TSS-C vs. MB (Tw-auto)	4.49×10^{-13}	$+14.77$
TSS-C vs. LLaMA-3 (Tw-auto)	5.15×10^{-4}	-2.83
TSS-BC vs. MB (Tw-gold)	1.81×10^{-4}	-2.54
TSS-BC vs. LLaMA-3 (Tw-gold)	5.11×10^{-13}	-12.47
TSS-BC vs. MB (Tw-auto)	4.32×10^{-3}	$+6.56$
TSS-BC vs. LLaMA-3 (Tw-auto)	$< 10^{-15}$	-10.64

Table 13: McNemar p and paired effect size d_z. MB = MentalBERT.

B.4 Paired Channel Contrasts

Dataset	Cmp	ΔF1	95% CI	p_{perm}
Dread (H)	C–A	$+0.106$	$[+0.07, +0.15]$	0.0001
Dread (H)	BC–A	$+0.057$	$[+0.02, +0.10]$	0.0032
Tw-gold (H)	C–A	$+0.039$	$[+0.02, +0.06]$	0.0005
Tw-gold (H)	BC–A	$+0.029$	$[+0.00, +0.05]$	0.0131
Tw-auto (A)	C–A	-0.019	$[-0.03, -0.00]$	0.9877
Tw-auto (A)	BC–A	-0.067	$[-0.08, -0.05]$	1.0000
Red-c. (A)	C–A	$+0.070$	$[+0.04, +0.10]$	0.0001
Red-c. (A)	BC–A	$+0.078$	$[+0.05, +0.11]$	0.0001

Table 14: Paired channel contrasts ($\Delta Macro - F1$) by dataset, with permutation p-values.

Channel/Model	Mean	Std	Min	Max
TSS-A	0.662	0.042	0.628	0.723
TSS-B	0.570	0.045	0.531	0.637
TSS-BC	0.679	0.032	0.653	0.734
TSS-C	0.685	0.015	0.666	0.700
MentalBERT	0.818	0.030	0.772	0.851

Table 15: Leave-one-domain-out cross-stressor robustness on Dreaddit (out-of-domain Macro-F1, mean over five domains).

Stressor	TSS-A	TSS-C	TSS-BC	MB
abuse	0.639	0.672	0.653	0.772
anxiety	0.723	0.700	0.734	0.828
financial	0.632	0.690	0.683	0.833
ptsd	0.686	0.695	0.669	0.851
social	0.628	0.666	0.659	0.806
Mean	0.662	0.685	0.679	0.818
Std	0.042	0.015	0.032	0.030

Table 16: Per-stressor LODO breakdown (MentalBERT = MB).

C Robustness Audits

C.1 LODO

C.2 Phenotyping

C.3 Channel Orthogonality

C.4 Efficiency

C.5 LLaMA-3 Few-Shot Instability

D Qualitative Conflict Zone

We extracted 800 instances (200 per dataset) for instance-level qualitative analysis of cases where the lexical baseline (Channel A) and the structural probes (Channels B/C) disagree. The full workbook (`error_analysis_qualitative.xlsx`) is released with the code.

Workbook statistics. The `Shift_Cases_AvsB` sheet documents 682 instances where Channel B corrects Channel A's errors: 415 (60.9%) are A-false-positives (keyword-driven false positives) and 267 (39.1%) are A-false-negatives (implicit stress). The complementary `Reverse_Shift` sheet documents 667 instances where A corrects B; restricting to human-labeled data yields 319 A-corrections (Twitter-gold: 200; Dreaddit-test: 119), comprising 177 B-false-positives and 142 B-false-negatives. The `All_Channels_Compare` sheet shows that Channel C alone corrects 253 of A's errors

Statistic	Value	95% CI / p
Bootstrap ARI	0.980	$[0.961, 0.991]$ (n_{boot}=100)
Cluster–dataset AMI	0.281	p=0.0005 (perm 2000)
Cluster–label AMI	0.084	p=0.0005 (perm 2000)

Table 17: K-Means (K=3) on Channel C (N=12,924, d=154). DoD analysis uses N=12,906; the 18-instance gap reflects zero-length POS sequences excluded from DoD but retained in clustering via Channel C features.

Summary	Value
Mean $\lvert\rho\rvert$ across pairs	0.595
Orthogonality ratio	0.095
A vs. B (key pair)	0.296
Effective N	271,026

Table 18: Instance-level Spearman ρ across channel pairs.

(195 keyword-driven false positives + 58 implicit stress), and BC corrects 241.

Table 21 gives 13 representative cases illustrating the two failure modes; the full characterization is in the released workbook.

E Extended Metric Dashboard

F SHAP Lexical Concentration

Channel	d	Train (s)	Reg.	ℓ_1-ratio
A	500	17.6	L_2	—
B	6	1007.9	ElasticNet	0.95
C	154	27.5	L_2	—
BC	160	335.9	ElasticNet	0.95

Table 19: Efficiency profile. Channel B's training time is an artifact of an exhaustive ElasticNetCV grid; once fitted, inference is near-instantaneous.

Dataset	F1	Prec.	Rec.	Pattern
Dreaddit-test (H)	0.613	0.606	0.984	Over-firing
Twitter-auto (A)	0.522	0.846	0.203	Under-firing
Twitter-gold (H)	0.800	0.780	0.680	Balanced

Table 20: LLaMA-3-8B (3-shot) precision–recall instability across label sources.

Case type	Sample (truncated)	GT	A	TSS
Keyword-driven FP (advert)	"call us at +971 4 324 3244 to consult … #MentalHealth #Stress #DryEye"	0	1	0
Implicit stress (suicidality)	"I am on the edge of committing… a few hours remaining…"	1	0	1
Boundary collapse (HR narrative)	"I man the front desk… HR Customer Service Representative…"	0	1	0
Keyword-driven FP (resilience)	"Resilience is not a trait… #mentalhealth #mentalstrength"	0	1	0
Keyword-driven FP (counseling ad)	"Sometimes love is complicated. Let's talk it out. #counseling #mentalhealth"	0	1	0
Keyword-driven FP (product ad)	"New Product Alert! Silicone Suction Snapper. Use code SOCIAL10 #Stress"	0	1	0
Keyword-driven FP (achievement)	"Squatting 315 lbs for 6 reps… Tonight I just did 305!!!! I'm so excited!!!"	0	1	0
Keyword-driven FP (policy)	"When moving into their tiny house, they would be given a state I.D.…"	0	1	0
Implicit stress (medical trauma)	"…they needed to strap me down… I completely dissociated."	1	0	1
Implicit stress (relational)	"The one person you thought you could talk to blows you off. tired alone hurt"	1	0	1
Implicit stress (masked depression)	"feel sad all the time then happy for a moment… faking it for days"	1	0	1
Implicit stress (burnout)	"stress at work… 60 hours a week (even when I had Covid)…"	1	0	1
Implicit stress (financial crisis)	"…paid 1,000 dollars for hospital bills… now we have nothing left."	1	0	1

Table 21: Representative qualitative conflict-zone cases. GT = ground truth, A = Channel A, TSS = Channel B/C combined verdict. The first eight cases are keyword-driven false positives where A fires on topical markers; the last five are implicit-stress cases where A misses distress expressed without explicit stress vocabulary.

Data	Ch.	F1	PR-AUC	MCC	B.Acc	Prev
Dread (H)	A	0.634	0.809	0.324	0.646	0.516
Dread (H)	C	0.739	0.823	0.495	0.740	0.516
Dread (H)	BC	0.690	0.819	0.421	0.697	0.516
Tw-G (H)	A	0.661	0.540	0.342	0.680	0.338
Tw-G (H)	C	0.701	0.637	0.422	0.689	0.338
Tw-G (H)	BC	0.690	0.708	0.438	0.677	0.338
Tw-A (A)	A	0.522	0.510	0.065	0.531	0.478
Tw-A (A)	C	0.504	0.656	0.209	0.568	0.478
Tw-A (A)	BC	0.455	0.658	0.167	0.545	0.478
Red-c (A)	A	0.611	0.951	0.223	0.608	0.880
Red-c (A)	C	0.681	0.964	0.371	0.710	0.880
Red-c (A)	BC	0.689	0.968	0.380	0.702	0.880

Table 22: Extended metric dashboard. Tw-G = Twitter-gold, Tw-A = Twitter-auto. Macro-F1 is the headline metric; PR-AUC is the imbalance-sensitive complement; MCC and balanced accuracy are further robustness checks under skewed prevalence.

Mode	n	Mean conc.	95% CI
lexical_wins	40	0.333	$[0.308, 0.355]$

Table 23: Top-10 SHAP concentration on Channel A (Lundberg and Lee, 2017) when lexical predictions beat structural ones. The top 10 features (2% of the 500-feature space) capture 33.3% of total absolute SHAP weight (16× the uniform-contribution expectation), consistent with strong shortcut reliance in SSI-rewarded regimes.

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

Mahule Roy[1] **Subhas Roy**[2]
[1]University of Oxford [2]TATA Consumer Products Limited
Correspondence: mroy25@bwh.harvard.edu

Abstract

Large language models achieve strong performance on biomedical question answering and summarization benchmarks, yet traditional evaluation metrics often fail to detect clinically significant factual errors. We introduce a unified evaluation framework that combines reference-based measures with evidence-grounded factuality verification to assess biomedical text generation. Evaluating four open-source models across three benchmarks (BioASQ, PubMedQA, MedLFQA), we find that 13.4–24.7% of generated claims are contradicted and 23–41% are unsupported, despite high lexical overlap scores. Our proposed Fact-Aligned Score (FAS) correlates strongly with claim-level verifiability ($\rho = 0.68$), substantially outperforming ROUGE-L ($\rho = 0.41$). We release an open-source toolkit with model outputs and analysis scripts to support reproducible factuality evaluation and safer deployment of biomedical LLMs.

1 Introduction

Large language models have shown remarkable capabilities in generating biomedical text for question answering, summarization, and clinical documentation. These models promise to assist clinicians, researchers, and patients by providing rapid access to structured and unstructured biomedical knowledge. However, current evaluation practices predominantly rely on lexical overlap metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and F1, which fail to capture clinically significant factual errors. A model may produce fluent and reference-aligned text while introducing subtle contradictions, unsupported claims, or outdated recommendations—errors that could have tangible consequences in healthcare settings. This gap creates an urgent need for evaluation frameworks that assess not only linguistic quality but also factual reliability and evidence alignment. In this work, we propose a unified evaluation framework that integrates traditional reference-based metrics with evidence-grounded factuality verification. Our framework systematically extracts atomic claims using spaCy (Honnibal et al., 2020), retrieves supporting biomedical literature from PubMed (Cock et al., 2009; Jin et al., 2019; Liu et al., 2005), and applies domain-adapted natural language inference (NLI) via BioBERT-NLI trained on MedNLI (Shivade, 2017; Gu et al., 2021) to verify each claim. By combining these components, we enable a more granular and clinically meaningful assessment of LLM outputs. This approach allows researchers and practitioners to identify factual inconsistencies, quantify reliability, and compare models not just on fluency, but on evidence-supported correctness, ultimately advancing the safe deployment of LLMs in biomedical contexts.

2 Related Work

Benchmark datasets such as BioASQ Task B (10th edition) (Nentidis et al., 2023), PubMedQA (Jin et al., 2019), and MedLFQA (Jeong et al., 2024) provide structured evaluation for biomedical QA across factoid, list, and long-form questions. BioASQ includes 492 test questions with 1–3 expert references, PubMedQA contains yes/no/maybe questions derived from PubMed abstracts, and MedLFQA provides 200 long-form QA pairs with clinician-verified claim-level annotations. While these resources are central to biomedical NLP evaluation, conventional metrics (e.g., ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019)) primarily capture lexical overlap and often overestimate performance by failing to detect unsupported or contradicted claims. Prior efforts have explored evidence-based and NLI-driven evaluation, including biomedical NLI models trained on MedNLI (Shivade, 2017). General-domain factuality metrics such as FactCC and QAFactEval

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 15–19
July 3-4, 2026 ©2026 Association for Computational Linguistics

focus on reference consistency within news or open-domain summarization, rather than external biomedical evidence alignment. In contrast, our framework integrates reference-based metrics with claim-level verification against retrieved PubMed abstracts, enabling detection of unsupported and contradicted scientific assertions even when reference answers are incomplete. By combining systematic multi-model evaluation with an open-source, reproducible toolkit, our approach advances clinically aligned factuality assessment for biomedical text generation.

3 Unified Evaluation Framework

3.1 Reference-based Accuracy

Reference-based accuracy measures the lexical and semantic alignment between generated and reference answers using ROUGE-1/2/L (F1), BLEU-4 (smoothing method 1), and BERTScore (DeBERTa-large-mnli, batch size 64, no baseline rescaling) via HuggingFace `evaluate`. For multi-reference datasets such as BioASQ, we adopt a maximum-score strategy following standard practice in BioASQ evaluations. These metrics provide a baseline measure of answer quality while capturing both n-gram overlap and contextual semantic similarity, forming the reference foundation for our composite factuality assessment.

3.2 Evidence-Grounded Factuality and Safety Assessment

Generated answers are decomposed into atomic claims using spaCy (Honnibal et al., 2020) and queried against PubMed (2014–2024) via Entrez API, retrieving the top 5 abstracts per claim. Each claim–abstract pair is evaluated with BioBERT-NLI (Shivade, 2017) (dmis-lab/biobert-large-nli-v1.1) using a 0.7 threshold to classify claims as supported, contradicted, or unsupported. Answer-level aggregation computes Support@5, Contra@5, and unsupported proportions, averaged per answer. Additionally, rule-based safety checks using RxNorm (Liu et al., 2005) and a curated FDA contraindication database (U.S. Food and Drug Administration, 2023) flag missing dosages or contraindicated medications. These safety flags are reported separately from the Fact-Aligned Score (FAS), a composite metric defined as $\text{FAS} = 0.3 \cdot \text{BERTScore} + 0.4 \cdot \text{Support@5} + 0.3 \cdot (1 - \text{Contra@5})$, which balances reference alignment, evidence support, and contradiction avoidance to provide a clinically meaningful measure of output reliability.

4 Experimental Setup

We evaluate three open-source generative models—Llama-2-13B-chat (Touvron et al., 2023), Llama-2-7B-chat (Touvron et al., 2023), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)—all of which are pre-trained and instruction-tuned but not further fine-tuned on the benchmark datasets. We also include PubMedBERT-base (Gu et al., 2021) as an encoder-only baseline. Evaluation is performed on three biomedical QA benchmarks: BioASQ Task B (10th edition; 492 test questions: 157 factoid, 138 list, 197 summary; average 1.4 references per question; mean reference length 87.3 tokens, SD 42.1) (Nentidis et al., 2023), PubMedQA (500 instances evenly split across yes/no/maybe; mean 42.6 tokens, SD 18.3) (Jin et al., 2019), and MedLFQA v1.0 (200-instance evaluation subset; long-form clinical QA pairs; claim-level factuality labels derived from the dataset's Must Have (MH) and Nice to Have (NH) annotations; mean 156.2 tokens, SD 67.8) (Jeong et al., 2024). Models were accessed via HuggingFace Transformers (v4.36.0) using the following identifiers: meta-llama/Llama-2-13b-chat-hf, meta-llama/Llama-2-7b-chat-hf, mistralai/Mistral-7B-Instruct-v0.2, and microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract (all accessed June 2024). Generation was performed with deterministic decoding (max_new_tokens=512, temperature=0, top_p=1, do_sample=False) in FP16 precision; PubMedBERT, being encoder-only, uses the source abstract as the answer for PubMedQA. Experiments were conducted on NVIDIA A100-40GB GPUs (156 total GPU-hours). Traditional metrics were computed using HuggingFace `evaluate` (v0.4.0): ROUGE-1/2/L (Lin, 2004) with Porter stemming and Penn Treebank tokenization (Bird et al., 2009), BLEU-4 (Papineni et al., 2002) with Chen and Cherry smoothing method 1, and BERTScore (Zhang et al., 2019) using `microsoft/deberta-large-mnli` (batch size 64). Factuality evaluation employed a modular pipeline with spaCy-based claim extraction (Honnibal et al., 2020), PubMed retrieval (2014–2024; top five abstracts per claim via Bio.Entrez), and NLI inference (batch size 16; 512-token truncation). Statistical significance was assessed using answer-level bootstrap resampling (10,000 iterations) (Efron and Tibshirani, 1986),

reporting two-tailed 95% confidence intervals; performance differences were considered significant when intervals excluded zero. Safety checks for clinical hazards leveraged the FDA contraindication database (U.S. Food and Drug Administration, 2023) and RxNorm (Liu et al., 2005).

For generation, we used the following prompt templates:

QA tasks (BioASQ factoid/list, PubMedQA):

```
Answer concisely and accurately based on
medical knowledge.
Question: {question}
Answer:
```

Summarization tasks (BioASQ summary, MedLFQA):

```
Provide a comprehensive summary based on
biomedical literature. Include findings,
mechanisms, and clinical implications.
Question: {question}
Summary:
```

5 Results

5.1 Traditional Metric Performance

All generative models evaluated (Llama-2-13B-chat, Llama-2-7B-chat, Mistral-7B-Instruct-v0.2) are pre-trained and instruction-tuned; no further fine-tuning was applied on the benchmark datasets. PubMedBERT is included as an encoder-only baseline. Table 1 reports reference-based performance across datasets. Llama-2-13B-chat achieves the highest ROUGE-L scores on all three benchmarks, with statistically significant improvements over Llama-2-7B on BioASQ and PubMedQA, though not on MedLFQA due to higher variance in long-form responses. Performance scales with model size, while Mistral-7B performs comparably to Llama-2-7B with no significant differences. PubMedBERT reveals a limitation of lexical metrics: on PubMedQA, it achieves ROUGE-L 0.412 despite not generating answers but simply reproducing the source abstract. This indicates that reference-based metrics reward lexical overlap even when outputs are not faithful answers to the question, motivating the need for evidence-grounded evaluation.

5.2 Evidence-Grounded Factuality

Table 2 presents Support@5 and Contra@5 scores. Although larger models demonstrate higher support rates, contradiction rates remain non-trivial (13.4%–24.7%), even for the strongest model. This confirms that high ROUGE does not imply factual reliability. Notably, PubMedBERT's relatively high ROUGE on PubMedQA (0.412) corresponds to lower support rates on BioASQ, illustrating dissociation between lexical similarity and claim-level verification.

Table 1: ROUGE-L scores (95% CI).

Model	BioASQ	PubMedQA	MedLFQA
Llama-2-13B	0.387±0.013	0.501±0.012	0.362±0.016
Llama-2-7B	0.358±0.013	0.473±0.012	0.334±0.016
Mistral-7B	0.365±0.013	0.481±0.012	0.341±0.016
PubMedBERT	0.187±0.013	0.412±0.012	0.143±0.016

Table 2: Factuality metrics (Support@5 and Contra@5) across BioASQ and PubMedQA. MedLFQA factuality is evaluated via the Fact-Aligned Score (FAS; Table 3). 95% CI shown.

Model	BioASQ S@5	BioASQ C@5	PubMedQA S@5
Llama-2-13B	71.2% [67.9-74.5]	13.4% [11.2-15.6]	78.4% [75.7-81.1]
Llama-2-7B	63.5% [60.2-66.8]	17.1% [14.9-19.3]	71.2% [68.5-73.9]
Mistral-7B	65.8% [62.5-69.1]	16.3% [14.1-18.5]	72.5% [69.8-75.2]
PubMedBERT	42.3% [39.0-45.6]	24.7% [22.5-26.9]	63.7% [61.0-66.4]

5.3 Per-Question-Type Analysis and Correlation with Gold Factuality

Factuality on BioASQ varies by question type: for Llama-2-13B, factoid questions (N=157) achieve the highest Support@5 (76.8%) and ROUGE-L (0.412) but also the highest contradiction rate (16.2%); list questions (N=138) show intermediate support (73.4%) and contradictions (14.1%), while summary questions (N=197) have lower support (65.2%) and fewer contradictions (10.8%). Differences between factoid and summary questions are statistically significant (Support@5 $\Delta = 11.6\%$, 95% CI [7.2, 15.9]; Contra@5 $\Delta = 5.4\%$, 95% CI [2.1, 8.7]). On MedLFQA, the Fact-Aligned Score (FAS) correlates most strongly with gold claim-level annotations ($\rho = 0.68$), outperforming ROUGE-L ($\rho = 0.41$) and BLEU-4 ($\rho = 0.36$); Support@5 alone also shows strong correlation ($\rho = 0.61$). Claim-level labels were derived from expert "Must Have" (MH) and "Nice to Have" (NH) statements, mapping MH to supported claims and treating unsupported or contradictory statements as factuality violations.

5.4 External Validation Considerations

Our primary validation relies on MedLFQA, from which we derive expert-informed claim-level fac-

tuality labels. While this enables direct quantitative correlation analysis, we acknowledge that additional human evaluation on BioASQ and PubMedQA would further strengthen external validity. To partially assess generalization, we analyzed consistency patterns across datasets. The relative ranking of models under FAS remained stable across BioASQ, PubMedQA, and MedLFQA, suggesting that the metric captures systematic reliability differences rather than dataset-specific artifacts. Table 3 shows the consistent ranking of models across all three datasets under our proposed FAS metric, with Llama-2-13B consistently outperforming other models and PubMedBERT serving as a lower-bound baseline. All evaluated models are pre-trained and instruction-tuned; no dataset-specific fine-tuning was applied. PubMedBERT is included as an encoder-only baseline.

Table 3: FAS scores (95% CI) across datasets. Higher is better. For MedLFQA, claim-level factuality labels were derived heuristically from expert-annotated "Must Have" (MH) and "Nice to Have" (NH) statements, mapping MH to supported claims and treating unsupported or contradictory statements as violations.

Model	BioASQ	PubMedQA	MedLFQA
Llama-2-13B	**0.612±0.013**	**0.703±0.012**	**0.584±0.016**
Mistral-7B	0.578±0.013	0.672±0.012	0.553±0.016
Llama-2-7B	0.571±0.013	0.665±0.012	0.547±0.016
PubMedBERT	0.412±0.013	0.589±0.012	0.398±0.016

5.5 Ablation Studies

We conduct ablation experiments on MedLFQA to assess the contribution of individual components of the FAS pipeline. Table 4 shows Spearman correlation with gold factuality under different configurations. Removing literature retrieval ("w/o retrieval") tests the importance of external evidence, resulting in a decrease from $\rho = 0.68$ to $\rho = 0.52$ ($\Delta = -0.16$). Replacing BioBERT-NLI with a general-domain MNLI model ("w/ general MNLI") evaluates the effect of domain adaptation, lowering correlation to $\rho = 0.45$ ($\Delta = -0.23$). Limiting retrieval to a single abstract per claim ("Single abstract") examines the impact of multi-document grounding, reducing correlation to $\rho = 0.56$ ($\Delta = -0.12$). These results suggest that both multi-document retrieval and domain-specific NLI significantly contribute to factuality alignment.

The NLI confidence threshold determines the minimum probability at which a claim–abstract pair is classified as supported or contradicted.

Table 4: Ablation on MedLFQA: Spearman ρ measures rank correlation between the Fact-Aligned Score (FAS) and gold claim-level factuality annotations. 95% CI shown; Δ indicates change relative to the full model.

Configuration	ρ	95% CI	Δ
Full model (FAS)	0.68	[0.62-0.74]	—
w/o retrieval	0.52	[0.45-0.59]	-0.16
w/ general MNLI	0.45	[0.38-0.52]	-0.23
Single abstract	0.56	[0.49-0.63]	-0.12

Claims with model confidence below the threshold are labeled as unsupported. To assess sensitivity, we varied this threshold from 0.5 to 0.9. Table 5 shows that Spearman correlation with gold factuality remains stable between 0.6–0.8, with optimal performance at 0.7, though differences within this range are not statistically significant. Based on these results, we selected 0.7 as the default threshold for our main analysis.

Table 5: Threshold sensitivity analysis on MedLFQA. Correlation with gold factuality at different NLI confidence thresholds.

Threshold	Spearman ρ [95% CI]
0.5	0.63 [0.56-0.70]
0.6	0.66 [0.59-0.73]
0.7	0.68 [0.62-0.74]
0.8	0.65 [0.58-0.72]
0.9	0.59 [0.52-0.66]

6 Conclusion

Open-source biomedical language models make 13–25% unsupported claims, with 4–7% flagged for missing dosages. Lexical metrics correlate weakly with factuality ($\rho = 0.36$–0.49), while the Fact-Aligned Score (FAS) aligns better ($\rho = 0.68$). Our framework—combining references, claim extraction, retrieval, and NLI—assesses biomedical text generation. Limitations include extraction errors, retrieval gaps, NLI mismatch, heuristic settings, English-only evidence (2014–2024), and multi-reference metric inflation.

References

Nentidis, A., Katsimpras, G., Krithara, A., Lima López, S., Farré-Maduell, E., Gasco, L., ... & Paliouras, G. (2023, September). Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical

semantic indexing and question answering. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 227-250). Cham: Springer Nature Switzerland.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., & Lu, X. (2019, November). Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 2567-2577).

Jeong, M., Hwang, H., Yoon, C., Lee, T., & Kang, J. (2024). Olaph: Improving factuality in biomedical long-form question answering. arXiv preprint arXiv:2405.12701.

Shivade, C. (2017). Mednli-a natural language inference dataset for the clinical domain.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Jiang, Y., Li, X., Zhu, G., Li, H., Deng, J., Han, K., ... & Zhang, R. (2023). 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. arXiv preprint arXiv:2311.09047.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python.

Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: prescription for electronic drug information exchange. IT professional, 7(5), 17-23.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

FDA, U. (2023). FDA adverse event reporting system (FAERS) public dashboard.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422.

Using Synthetic Records to Improve Automated Identification of Seizure Freedom in Clinical Text about People with Epilepsy

Stephen H. Barlow[1], Yujian Gan[1,2], Joe Davies[1], Joel S. Winston[1], James T. Teo[1,3], Mark P. Richardson[1], Ben Holgate[1],
[1]Department of Basic & Clinical Neuroscience, King's College London, United Kingdom
[2]School of Electronics, Electrical Engineering & Computer Science, Queen's University Belfast, United Kingdom
[3]Neurosciences Institute, Cleveland Clinic London, United Kingdom
Correspondence: benjamin.holgate@kcl.ac.uk

Abstract

Seizure freedom is a key clinical outcome for people with epilepsy (PWE) yet it is primarily recorded in free-text notes and letters in the United Kingdom, making it difficult to aggregate and track at scale. This paper introduces a generative LLM-based pipeline boosted by synthetic data to identify a PWE's seizure freedom status in clinicians' records. We fine-tuned seven different large language models (LLMs) with between 4-14 billion parameters using LoRA to compare models trained on synthetic records against those trained on expert annotated records. The best performing configuration, based on Qwen-2.5-14B, was trained entirely on synthetic records and used chain-of-thought (CoT) reasoning (both generated by GPT-5). This achieved an F1 score of 0.90±0.02 on double-annotated test data and outperformed the equivalent model trained on authentic clinician records, which achieved 0.87±0.04. The synthetically trained models also have the benefit of outputting their CoT reasoning process for greater decision-making transparency and can also make use of unused supervised training data for significantly increased test examples. This work has implications for monitoring a key treatment outcome for PWE automatically and at scale.

1 Introduction

Epilepsy is a chronic neurological condition characterised by recurrent seizures (Beghi, 2019). Seizure freedom is a key outcome for improving quality of life in people with epilepsy (PWE) (Poochikian-Sarkissian et al., 2008) and an important post-surgical outcome (Wieser et al., 2001). However, its definition remains inconsistent (Halford and Edwards, 2020). Engel's (1993) post-operative classification prioritises reduction of "disabling seizures", placing less emphasis on minor seizures, while Wieser et al. (2001) proposed a stricter definition requiring complete seizure freedom without auras, aligning more closely with PWE's expectations. Clinically, seizure freedom also requires a temporal threshold, as PWE may experience seizure recurrence after years without them. For example, the UK Driving and Vehicle Licensing Agency (DVLA) requires 12 months of seizure freedom for licence reinstatement (DVLA, 2026). Other studies have used six- and twelve-month thresholds (Hsieh et al., 2023; Choi et al., 2014). Thus, seizure freedom is both a definitional and temporal concept.

In the UK, difficult-to-treat epilepsy is managed in outpatient (ambulatory) settings, where consultation summaries (outpatient letters) are recorded as unstructured or semi-structured free text. This format limits automated information extraction. Advances in natural language processing (NLP), particularly large language models (LLMs), have enabled extraction of complex contextual information from electronic health records (EHRs) (Jerfy et al., 2024; Jin et al., 2025), including temporality (Yuan et al., 2024), which is essential for determining seizure freedom.

Accordingly, this study develops an LLM-based seizure-freedom classifier trained entirely on synthetic epilepsy letters which provides generated explanations. We also examine how training data origin (synthetic vs. authentic), class distribution, and synthetic chain-of-thought reasoning affect model performance. While this paper explores an epilepsy-specific application, we believe this approach has the potential to improve other imbalanced clinical information extraction tasks.

2 Related Work

Previous studies have applied NLP to extract seizure freedom-status from epilepsy records. Xie et al. (2022) used BERT-style models (Devlin et al., 2019) to extract seizure freedom and frequency, defining seizure freedom as no seizures since the last clinic visit or for at least 12 months.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 20–30
July 3-4, 2026 ©2026 Association for Computational Linguistics

Due to model input length limits, they used paragraph-level classification. Fernandes et al. (2024) found performance declined without reliable paragraph selection, and later work by Xie et al. (2023) introduced a bespoke paragraph extractor to aggregate document-level predictions. However, seizure freedom is inherently a document-level construct, true or false at the time of writing. Modern generative LLMs with longer context windows could perform direct document-level classification without rule-based passage selection, which could in turn better resolve contradictory statements. Related work on seizure frequency also reflects these broader NLP developments, progressing from rule-based methods (Fonferko-Shadrach et al., 2019) to encoder-only models (Xie et al., 2022), in-context learning (Holgate et al., 2024), and parameter-efficient fine-tuning for document-level approaches (Holgate et al., 2025).

These studies usually rely on relatively small annotated datasets, reflecting expert time constraints and limiting representation of rarer contexts. Data imbalance is intrinsic, as seizure-free PWE attend fewer appointments. Synthetic data offers a potential solution, enabling larger training datasets while reserving expert-annotated data for evaluation. Teacher–student approaches, where large LLMs generate synthetic training data for smaller models, are a potential solution. Phi-4 (Abdin et al., 2024), for example, was trained primarily on synthetic GPT-4 output. Synthetic data has also addressed healthcare challenges like Šuvalov et al. (2025) creating an Estonian clinical named entity recognition system using GPT-generated records. However, it remains unclear whether fully synthetic training can capture temporal clinical concepts like seizure freedom. Additionally, approaches such as DeepSeek's synthetic reasoning chains (Guo et al., 2025; Liu et al., 2025) may enhance transparency by addressing clinician concerns about "black box" decisions (Salvi et al., 2025) by providing interpretable reasoning in model output.

Synthetic clinical text has also been explored in epilepsy where Goldenholtz et al. (2025) used LLM-generated notes to simulate a clinical trial. This study was entirely *in silico* and thus highlights the need to evaluate synthetic training against real-world clinician-authored notes, which this study addresses.

3 Methods

Seizure freedom (SF) classification is framed in this work as a binary classification task using generative language modelling. We compare authentic and synthetic training datasets to examine the effect on performance for classifying seizure-freedom. For the synthetic training approach, we also compare how additional fine-tuning on CoT reasoning-chains affects performance and outputs. For the authentic letters we also tested whether under sampling the majority class resulted in better performance due to the pronounced class imbalance in this task.

3.1 Seizure Freedom Definition

Seizure freedom is considered an important outcome of epilepsy treatment, yet definitions differ. We sought to find a compromise between a the stricter 12-month definition of seizure freedom (DVLA, 2026; Choi et al., 2014) and Xie et al.'s (2022) last clinic approach. Hsieh et al.'s (2023) six-month definition of seizure freedom provided a clinically relevant outcome while providing enough training and testing examples (as 12-months seizure free was much less frequently seen in our dataset than six-months). Accordingly, in this work seizure freedom is defined as:

> *"The person with epilepsy has suffered no epileptic seizures of any kind for a period of six months or more."*

Figure 1 shows some artificial examples of ambiguous examples of SF-positive and SF-negative letters by this definition. These characteristics informed the need to use LLMs with longer context windows (the number of text tokens an LLM can process at a time) as relevant information could be in any part of (sometimes very long) letters, and a degree of temporal reasoning is required to ascertain the correct SF-status.

3.2 Authentic Dataset

We randomly sampled and annotated for SF status 2,114 epilepsy letters from King's College Hospital (London, UK) with a team of five epilepsy experts. These were retrieved using the information retrieval system CogStack (Jackson et al., 2018). The letters were written between 1 January 2013 and 30 September 2023. The documents comprise doctor's and nurse's reports with the vast majority referring to outpatient clinic visits. A

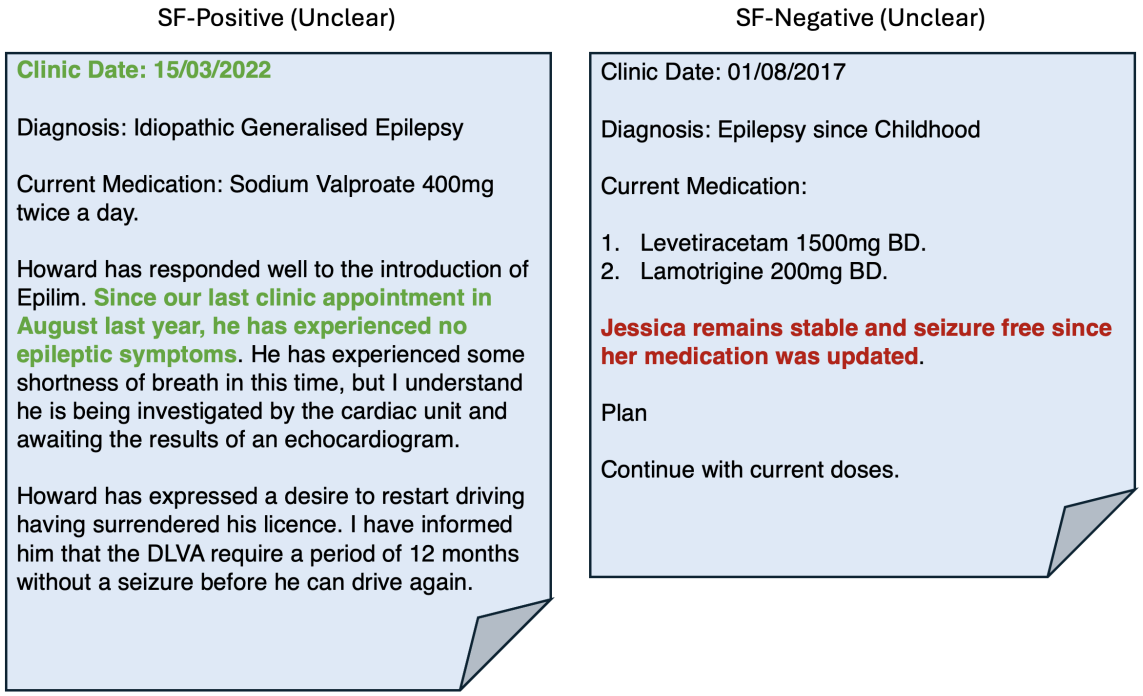


Figure 1: Fictional clinician letters for epilepsy demonstrating ambiguous SF-positive and SF-negative status. Letters can be much longer and more detailed in practice. Both examples show how temporal reasoning is required to make the correct classification (either relating information to provided dates, or recognising there is missing temporal information). The negative example demonstrates how we consider all examples that are not explicitly seizure free (for six months or more) as negative examples.

subset of 300 reports were dual annotated independently by epileptologists to test inter-annotator agreement and subsequently a gold-standard test set was produced via consensus between the experts. The overall agreement for seizure freedom was 0.85 using Cohen's Kappa, representing 'near perfect' (Cohen, 1960) or 'strong' (McHugh, 2012) agreement. SF-positive status was observed to be rare in the hospital letters, around 6% of the total. Therefore, due to limited resources for annotation, the bulk of the letters were annotated by one person. This motivated both the creation of the dual-annotated subset (for greater confidence in evaluation) and the need to create synthetic letters for training.

This project and its data usage operated under UK Health Research Authority (HRA) London South East Research Ethics Committee approval (reference18/LO/2048 and renewed 24/LO/0057) granted to the King's Electronic Records Research Interface (KERRI) with data research opt-out. Individual consent for participation was waived by the KERRI committee at King's College Hospital (KCH) for purposes of evaluating NLP for Epilepsy (approved February 2021) with local institutional oversight.

3.3 Synthetic Dataset

The aforementioned class imbalance motivated the use of synthetic training examples. As part of a wider synthetic data project 15,099 synthetic epilepsy letters were generated by GPT-5 to reflect a range of different seizure frequency and seizure freedom occurrences in clinical epilepsy letters (Gan et al., 2026). These were created from 10 manually composed letters (to mimic a UK style of epilepsy reporting) and 74 templates of different letter characteristics (different seizure frequencies, no seizures mentioned, seizure free etc.). Once these letters were generated, they were fed back to GPT-5 where it was prompted to ascertain the correct seizure frequency/seizure freedom-status and provide a description of how it came to that conclusion. This functions as a quality-assurance step for the synthetic data as any letters that GPT-5 does not correctly categorise were discarded. From these explanations 454 synthetic letters were found that were correctly described by GPT-5 as implying SF-positive status. These were extracted along with 454 SF-negative examples to create a balanced synthetic training dataset. This approach mitigates the class imbalance issue present in a true clinical distribution of

letters. Two variations on this synthetic dataset were trialled: Synthetic and Synthetic-CoT. For the latter we include the reasoning chains generated by GPT-5 as part of the output for supervised training to test if it improves model performance. This also has the benefit of providing text explanations for the classifications, something that could potentially assist clinicians (Miao et al., 2024). This project respects OpenAI's Terms of Use. This study is conducted for academic, non-commercial research purposes. We do not release distilled model weights and do not position the distilled model as a deployable substitute for OpenAI services; results are reported for methodological analysis.

3.4 Dataset Splits, Evaluation, and Prompting

We define four training datasets: Authentic, Authentic-Balanced, Synthetic and Synthetic-CoT. The Authentic training set uses 1,814 annotated letters (all but the 300 double-annotated letters). The Authentic-Balanced dataset uses the 115 SF-positive examples that are not included in the gold-standard test set alongside a matching 115 SF-negative examples. This matches the distribution of the Synthetic and Synthetic-CoT training datasets which contain 454 SF-positive and 454 SF-negative examples.

We used two test datasets. The gold-standard test set consisted of the 300 examples with consensus annotation from two experts and both authentic and synthetic models used this for testing. The silver-standard test set consisted of all 2,114 annotated letters and was used for additional testing of the synthetically trained models. This allowed for a better indication of long-term performance. The authentic models were not tested on this as they were explicitly trained on all but the 300 gold-standard examples in this dataset.

As seizure freedom extraction is framed as a binary classification task, we used precision, recall, F1 score, and accuracy for evaluation. Due to the pronounced class imbalance, we prioritised macro-averaged F1 over accuracy to compare the different models. This ensured we evaluated both the majority and minority classes equally and avoided making decisions due to deceptively high accuracy scores caused by the class imbalance.

The prompt used for training and inference is demonstrated in Figure 2. The additional supervision signal of GPT-5 reasoning chains were prepended to the desired output in the Synthetic-CoT dataset. Figure 2 also demonstrates how models trained on the Synthetic-CoT dataset generate temporal reasoning before making a final classification decision. The other training approaches used the same input prompt but only provided a 'Seizure free: yes/no' output.

Base Model	Parameters	Type
Gemma-3	4B	VLM
MedGemma	4B	VLM
Qwen-2.5-7B	7B	LLM
Lingshu	7B	VLM
Llama-3.1-Instruct	8B	LLM
Ministral-Instruct-2410	8B	LLM
Qwen-2.5-14B	14B	LLM

Table 1: Comparison of the LLMs used in this study. 'B' represents 'Billion'.

3.5 LLMs and Training Approach

We selected seven open-weight LLMs with between 4 billion to 14 billion parameters for fine-tuning on the SF extraction task (Table 1). Some of these were vision-language models (VLMs), but only their language capabilities were utilised for this task. The selection represents both general and clinically specialised models and were chosen as they represented a diverse selection of different size LLMs with different pre-training corpora and model architectures. We also trained two simpler, logistic regression classifiers using term frequency-inverse document frequency (TF-IDF) encodings to serve as baselines, one on the 'Authentic' dataset, and one on the balanced 'Synthetic' dataset.

Early experiments revealed that in-context learning was not sufficient for adequate performance on this task so fine-tuning was performed using low-rank adaptation (LoRA) (Hu et al., 2021). Due to the extra VRAM (Video Random Access Memory) Qwen-2.5-14B required for fine-tuning, we used 8-bit quantization using Bitsandbytes (Dettmers et al., 2021). The optimisation task was autoregressive language modelling, where the model is trained to predict the next token of text from previous text. The appropriate instruction templates and tokenizers were used for each model tested.

No specific hyperparameter tuning took place in order to compare synthetic and authentic approaches without biasing training towards either

Chain-of-Thought Fine-Tuning Prompt

Input:

"You are a professional neuroscientist. The following is a document written by a healthcare professional. Analyze the text and ascertain whether the patient referred to in the letter has been seizure free for six months or more. Think through the problem step-by-step. Some things to consider include: Ascertain if the letter contains information about seizures, ascertaining if they have had recent seizures, and if they have not had seizures for six months or more. Finally, if you think the subject has been seizure free for six months or more return 'Seizure free: yes', otherwise return 'Seizure free: no'.

Document:

[LETTER]"

Output:

"The letter explicitly states: "She reports ongoing absence of seizures since February 2023." The letter date is 02 October 2025. Calculate duration of seizure freedom from Feb 2023 to Oct 2025: from Feb 2023 to Feb 2024 = 12 months; Feb 2024 to Feb 2025 = 12 months; Feb 2025 to Oct 2025 = 8 months. Total = 32 months. This exceeds 6 months, so she is seizure free, and the duration is multiple years. The allowed format for multi-year seizure freedom is 'seizure free for multiple year' when exact number in years isn't required. ***Seizure free: yes"***

Figure 2: The main prompt used for the Synthetic-CoT training strategy. The Authentic, Authentic-Balanced and Synthetic training datasets used the same prompt but "Seizure free: yes/no" only as the desired output. [LETTER] signifies where the authentic or synthetic letter would be inserted into the prompt.

one. All models were trained for three epochs with a batch size of four. We used a learning rate of 0.0001 with a cosine schedule and a warmup ratio of 0.1. LoRA rank, alpha and dropout were set at 8, 16 and 0 respectively. Greedy decoding was used for text generation to ensure a deterministic answer. For synthetic data generation GPT-5's temperature was set (via API) to 0.

4 Results

Table 2 shows the distribution of SF-positive and SF-negative examples in the different datasets used for training and testing. The extracted clinical data (Authentic, Authentic-Silver and Authentic-Gold) demonstrated a pronounced class imbalance that is rectified by under sampling in the Authentic-Balanced, Synthetic, and Synthetic-CoT training datasets. There were only 21 SF-positive letters in the gold standard dataset but this is increased to 136 in the silver standard test set, demonstrating the benefit of having additional annotated data available for testing.

Figure 3 shows how synthetic and authentic training datasets affect performance on the gold-standard test set. Synthetic training tended to perform as well, if not better, than authentic training. For authentic training, using a balanced dataset did not help and, in some cases, seemed to hinder the model. The use of CoT reasoning chains in training improved performance on certain models but not others. All the LLM-based approaches dramatically outperformed the TF-IDF/logistic regression model which only achieved an F1 score of 0.48 (in contrast to F1 scores of 0.74-0.91 for the LLM-based pipelines).

Figure 4 explores the performance of synthetically trained models further on the larger silver-standard test set. This provides a more thorough demonstration of synthetically trained model performance with 2,114 test examples as compared to 300 (Table 2). It was observed again that benefits for training with CoT reasoning chains were dependent on the individual base LLM. The Qwen models seemed to find the additional CoT reasoning supervision beneficial on the larger test dataset with Qwen-2.5-14B model's F1 score improving to 0.85 from 0.84, and Qwen-2.5-7B improving to 0.84 from 0.82. The best performing model was Qwen-2.5 14B with CoT reasoning and this was also a strong performer on the gold standard subset with an F1 score of 0.88 (Figure 3). In

Dataset	Use	No. Letters	SF-Positive	SF-Negative
Authentic	Training	1814	115	1699
Authentic-Balanced	Training	230	115	115
Synthetic	Training	908	454	454
Synthetic-CoT	Training	908	454	454
Authentic-Silver	Synthetic Test	2114	136	1978
Authentic-Gold	Synth and Authentic Test	300	21	279

Table 2: The different training and test sets used in this work and their distribution of SF-positive and negative examples.

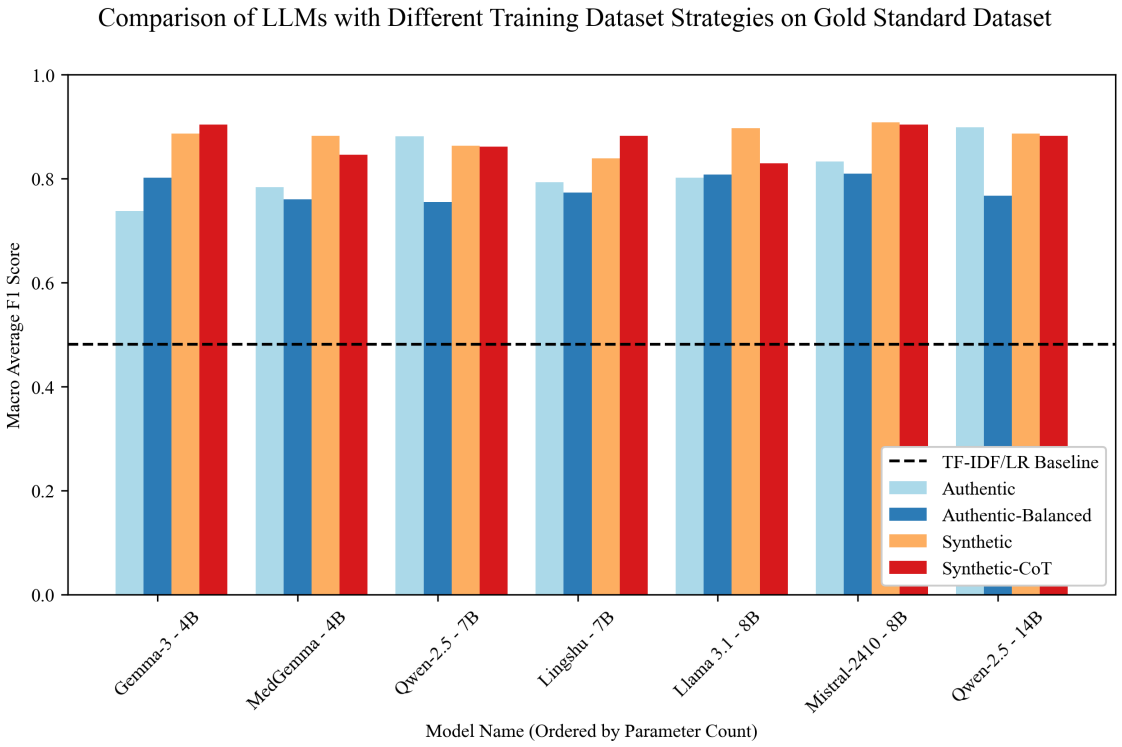


Figure 3: Comparison of the different LLMs trialled when trained on four different dataset strategies. Results are on the gold-standard test set. A TF-IDF logistic regression classifier (trained on the authentic training dataset) is provided as a baseline.

contrast, LLama 3.1-8B performed better without CoT reasoning supervision with its F1 score performance on the silver standard test set dropping to 0.77 from 0.83. The synthetically-trained TF-IDF/logistic regression baseline performs poorly and only achieves an F1 score of 0.24.

As Qwen-2.5-14B was the best performing LLM, when taking into account results on both the gold and silver test sets, Table 3 shows its performance in more detail when trained three times with different random seeds. The Synthetic-CoT training set resulted in the best performing model pipeline on both the gold and silver-standard datasets with F1 scores of 0.90±0.02 and 0.85±0.02 respectively.

5 Discussion

Our best model configuration for classifying seizure freedom status, using Qwen-2.5-14B and the Synthetic-CoT training strategy, achieved F1 scores of 0.90±0.02 on the gold standard test and 0.85±0.02 on the larger silver standard test dataset. The strong performance on both suggests it will be robust to different examples when used in practice. This model was trained entirely on synthetic data using a CoT reasoning approach for training. This means it generates the LLM-reasoning steps the model took to come to a final answer as part of the output. It is difficult to directly compare our results with Xie et al. (2022; 2023) and Fernandes et al.'s (2024) approaches to seizure freedom classification as the models were trained on different datasets and use different definitions of seizure freedom. However, our best model configuration compares favourably to the 0.83-0.88 F1 scores reported by Xie et al. (2022; 2023) (who also had a gold-standard test set of 300 examples) and 0.38 reported by Fernandes et al. (2024). The lower performance of the latter was attributed to using Xie et al.'s (2022) model without fine-tuning on their own letters. This demonstrates the issue of model generalization to external data. The goal of our research was to maximise performance for a stringent definition of seizure freedom rather than tackle broader seizure frequency concerns. Seizure freedom is an important outcome in epilepsy treatment so ensuring that we automatically identify PWE who have had at least six months of seizure freedom at a point in time was the priority. Although we explore a specific clinical use case in this paper, we believe the results of these experiments are also evidence to support using synthetic training data for other medical NLP tasks.

A key finding from this work is that a successful seizure freedom classifier can be trained using only synthetically generated training data. This provides two key benefits. It allows all expert annotated examples to be used for testing, and reduces privacy concerns by not using any sensitive data for fine-tuning. The latter makes it impossible for the model to disclose this data from leakage (Zhong et al., 2025), or adversarial attack (Rahman, 2023). Our best overall model fine-tuned Qwen-2.5-14B on synthetic letters and CoT reasoning. The Qwen 2.5 models were trained with an emphasis on reasoning (Yang et al., 2024), which is a potential reason they benefited from the CoT training signal, particularly on the silver-standard test set (Figure 4). Llama 3.1, in contrast, was not specifically optimised for reasoning which could explain why performance worsened for this LLM using CoT training signal. We identify two potential reasons for how a synthetically trained model can outperform an authentically trained one. Firstly, the increased training examples that the synthetic approach allows provides greater supervision signal. Secondly, there are likely less ambiguous synthetic examples (as these have been verified by GPT-5) so the model is less likely to overfit to them. These factors could also interact, as an ambiguous authentic example will have a disproportionately large effect due to the smaller number of training examples overall.

Synthetic data has been proposed as a solution to data bottlenecks when training healthcare models (McDuff et al., 2023). However, it is non-trivial creating examples which replicate authentic documents faithfully (Smolyak et al., 2024). Epilepsy letters are a good case study for synthetic data approaches as concepts like seizure freedom and frequency require dates (clinic dates, medication changes, etc.) to remain in letters for training. This is because dates and sentences referring to them may be the only reference to seizure status in the text. This can be considered private information and some de-identification models will remove them by default. Synthetic examples overcome this limitation by allowing student models to train on temporal relationships without jeopardising patient privacy. The reasoning chain demonstrated in Figure 2 shows how both the teacher and student models make use of this

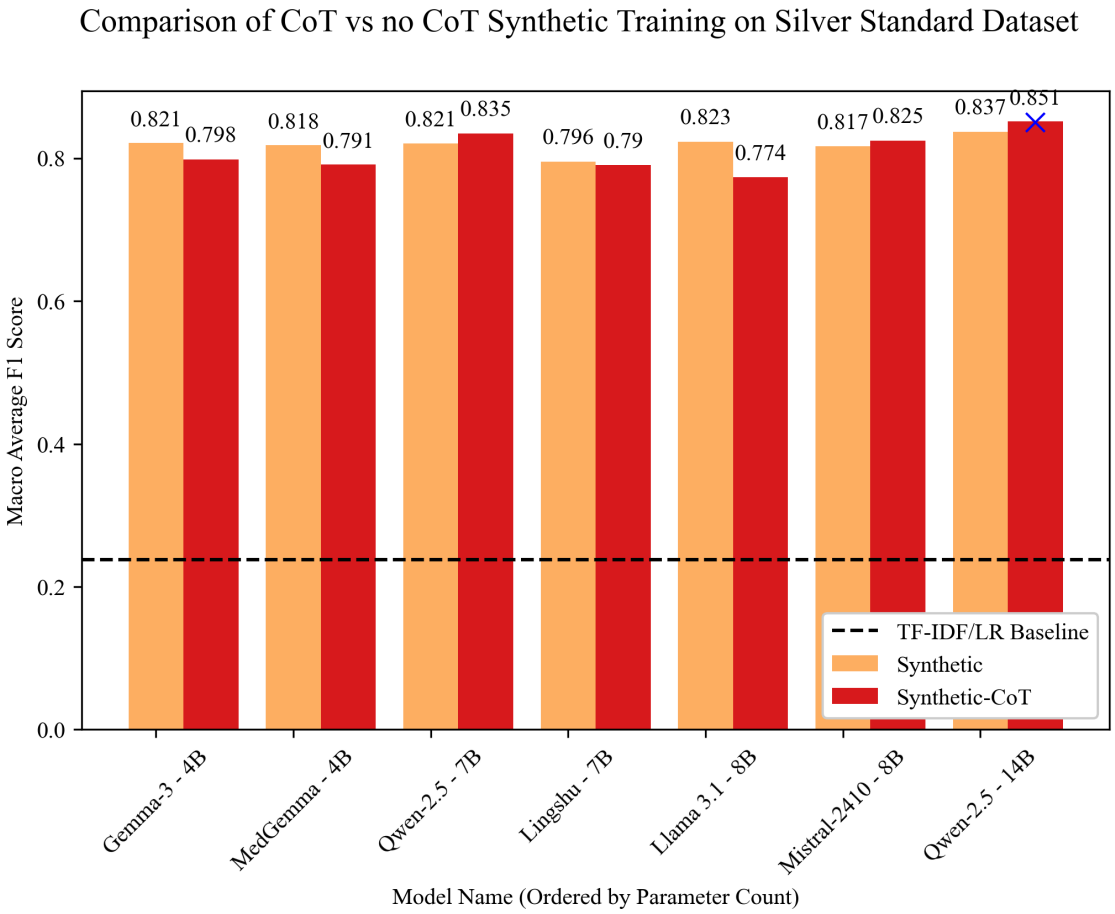


Figure 4: Comparison of Synthetic and Synthetic-CoT training strategies on the silver-standard test set. The blue cross represents the best performing model by macro average F1 score. Exact F1-score values are stated above the bars. A TF-IDF logistic regression classifier (trained on the synthetic training dataset) is provided as a baseline.

Training Set	Test Set	Accuracy	F1	Precision	Recall	Support
Authentic	Gold	0.972±0.009	0.874±0.049	**0.952±0.011**	0.824±0.062	300
Auth-Balanced	Gold	0.926±0.015	0.793±0.026	0.739±0.026	**0.909±0.012**	300
Synthetic	Gold	0.969±0.002	0.88±0.007	0.883±0.025	0.881±0.028	300
Synth-CoT	Gold	**0.973±0.007**	**0.896±0.023**	0.904±0.037	0.89±0.016	300
Synthetic	Silver	0.954±0.007	0.829±0.014	0.799±0.027	0.87±0.013	2114
Synth-CoT	Silver	***0.961±0.004***	***0.847±0.019***	***0.826±0.016***	***0.873±0.03***	2114

Table 3: Performance of Qwen-2.5-14B when LoRA fine-tuned with the four different training strategies on both test sets. Each model was trained three time with three random seeds, and the mean and standard deviation is reported. F1, precision, and recall metrics are macro-averaged. Support signifies the number of examples in the test sets. Bold text represent the best performing on the gold-standard test set, and bold plus italics represents the best performing on the silver-standard test set.

temporal information to arrive at a classification. By using a high-quality reasoning model as the teacher model, like GPT-5, we demonstrate that student models can learn to generate reasoning chains that justify the classification of SF-status. This can not only improve performance of the model (depending on the student LLM used), but also provides a human interpretable explanation for the clinician, something that has been identified as important step towards building confidence in AI solutions for healthcare (Miao et al., 2024). The reasoning chains also provided another benefit by serving as a verifier for quality when creating the synthetic letters. This combination of synthetic letters with CoT reasoning appears to be a powerful combination for developing epilepsy-specific systems where temporal features are directly related to patient outcomes.

The synthetic training data approach also has advantages when it comes to evaluation. Many classification models are tested on a small number of examples (Hou et al., 2024), which reduces the support for any classification metrics presented. Not requiring annotations for training means all annotations can go towards verifying the performance of models. The synthetically trained models in this project were evaluated on seven times the number of samples than was possible for the authentically trained models, providing greater confidence in them.

Although the largest model tested was the most effective, it should be noted that all the models tested reliably extracted SF-status from reports (Figures 3 and 4). Many hospitals do not have extensive computational resources, so having a smaller, 4 billion parameter LLM as a viable option is encouraging for clinical translation. It is also worth noting that the smaller models seemed to particularly benefit from the additional training examples provided by the synthetic letters. This is illustrated in Figure 3 where Gemma and MedGemma performed comparatively poorly when trained on authentic letters compared to synthetic letters (on the gold-standard test set). As multimodal AI becomes more prevalent, it was also encouraging to see that the VLMs (Gemma, MedGemma, and Lingshu) did not perform noticeably worse than LLMs when operating using text only. This suggests there is little drawback in training foundation models for multimodality outside of data resource and compute constraints.

For future work, it would be useful to apply synthetic-only training approaches to other clinical NLP tasks. Identifying people with rare phenotypes (for both epilepsy and other conditions) in electronic health records would be a valuable test for this approach. We also plan to implement the best performing model from this project as a feature extractor for an epilepsy prognostic tool, where a PWE's seizure freedom status is likely to have predictive value.

6 Conclusion

We developed an LLM-based pipeline to classify PWE's clinical reports for seizure freedom status. A configuration using Qwen-2.5-14B trained on a dataset consisting entirely of synthetically generated clinician letters from GPT-5 with CoT reasoning chains performed the best. This model offered key advantages over the traditional supervised approach. Most notably the CoT from the teacher model, GPT-5, can be used as supervision signal and allow the student model, to generate CoT text for better performance and greater transparency in its answers. Also, not needing annotated data for training allowed the model to be evaluated on a much larger test set. This pipeline has the potential to be used for monitoring the progress of PWE at scale and determining if medication changes or other interventions result in this key clinical outcome being achieved.

Limitations

There were three main limitations. Firstly, we were unable to perform external testing for the authentic models due to only having annotations from one hospital. Secondly, the silver-standard test set was only single-annotated and although we found these annotations to be reliable, there could be more annotation mistakes in this test set when compared to the gold-standard test set. Finally, these are document-level classification models, and there are cases where relevant temporal information could be spread across multiple documents which this method will not take into account.

Acknowledgements

This document was produced as part of a research project funded by the Epilepsy Research Institute (grant ref 2209) and by an investigator-initiated research grant from Angelini Pharma.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, and Piero Kauffmann. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Ettore Beghi. 2019. The epidemiology of epilepsy. *Neuroepidemiology*, 54(2):185–191.

Hyunmi Choi, Marla J. Hamberger, Heidi Munger Clary, Rebecca Loeb, Frankline M. Onchiri, Gus Baker, W. Allen Hauser, and John B. Wong. 2014. Seizure frequency and patient-centered outcome assessment in epilepsy. *Epilepsia*, 55(8):1205–1212.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *ArXiv*, abs/2110.02861.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

DVLA. 2026. Epilepsy and driving. `https://www.gov.uk/epilepsy-and-driving` [Accessed: 2026-02-27].

Jerome Engel, Jr., P.C. Van Ness, T. Rasmussen, and L.M. Ojemann. 1993. Outcome with respect to epileptic seizures. In *Engel, J. Jr (Ed.), Surgical Treatment of the Epilepsies*, pages 609–621. Raven Press, New York.

Marta Fernandes, Aidan Cardall, Lidia M. V. R. Moura, Christopher McGraw, Sahar F. Zafar, and M. Brandon Westover. 2024. Extracting seizure control metrics from clinic notes of patients with epilepsy: A natural language processing approach. *Epilepsy Research*, 207:107451.

Beata Fonferko-Shadrach, Arron S. Lacey, Angus Roberts, Ashley Akbari, Simon Thompson, David V. Ford, Ronan A. Lyons, Mark I. Rees, and William Owen Pickrell. 2019. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the exect (extraction of epilepsy clinical text) system. *BMJ Open*, 9(4):e023232.

Yujian Gan, Stephen H. Barlow, Ben Holgate, Joe Davies, James T. Teo, Joel S. Winston, and Mark P. Richardson. 2026. Reproducible synthetic clinical letters for seizure frequency information extraction. *Preprint*, arXiv:2603.11407.

Daniel M. Goldenholz, Shira R. Goldenholz, Sara Habib, and M. Brandon Westover. 2025. Inductive reasoning with large language models: A simulated randomized controlled trial for epilepsy. *Epilepsy Research*, 211:107532.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and Xiao Bi. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jonathan J. Halford and Jonathan C. Edwards. 2020. Seizure freedom as an outcome in epilepsy treatment clinical trials. *Acta Neurologica Scandinavica*, 142(2):91–107.

Ben Holgate, Joe Davies, Shichao Fang, Joel Winston, James Teo, and Mark Richardson. 2025. Fine-tuning llms to extract epilepsy seizure frequency data from health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 44–55.

Ben Holgate, Shichao Fang, Anthony Shek, Matthew McWilliam, Pedro Viana, Joel S Winston, James T Teo, and Mark P Richardson. 2024. Extracting epilepsy patient data with llama 2. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 526–535.

Rui Hou, Joseph Y. Lo, Jeffrey R. Marks, E. Shelley Hwang, and Lars J. Grimm. 2024. Classification performance bias between training and test sets in a limited mammography dataset. *PLOS ONE*, 19(2):e0282402.

Jason K. Hsieh, Francesco G. Pucci, Swetha J. Sundar, Efstathios Kondylis, Akshay Sharma, Shehryar R. Sheikh, Deborah Vegh, Ahsan N. Moosa, Ajay Gupta, Imad Najm, Richard Rammo, William Bingaman, and Lara Jehi. 2023. Beyond seizure freedom: Dissecting long-term seizure control after surgical resection for drug-resistant epilepsy. *Epilepsia*, 64(1):103–113.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. 2018. Cogstack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC Med Inform Decis Mak*, 18(1):47.

Aadit Jerfy, Owen Selden, and Rajesh Balkrishnan. 2024. The growing impact of natural language processing in healthcare and public health. *INQUIRY:*

The Journal of Health Care Organization, Provision, and Financing, 61:00469580241290095.

Myeong Jin, Sang-Min Choi, and Gun-Woo Kim. 2025. Comcare: A collaborative ensemble framework for context-aware medical named entity recognition and relation extraction. *Electronics*, 14(2).

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, and Chen Dong. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Daniel McDuff, Theodore Curran, and Achuta Kadambi. 2023. Synthetic data in healthcare. *arXiv preprint arXiv:2304.03243*.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–82.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148.

Sonia Poochikian-Sarkissian, Souraya Sidani, Richard Wennberg, and Gerald M. Devins. 2008. Seizure freedom reduces illness intrusiveness and improves quality of life in epilepsy. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, 35(3):280–286.

Mohamed. A. Rahman. 2023. A survey on security and privacy of multimodal llms - connected healthcare perspective. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 1807–1812.

Massimo Salvi, Silvia Seoni, Andrea Campagner, Arkadiusz Gertych, U. Rajendra Acharya, Filippo Molinari, and Federico Cabitza. 2025. Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *International Journal of Medical Informatics*, 197:105846.

Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA open*, 7(4):ooae114.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using synthetic health care data to leverage large language models for named entity recognition: Development and validation study. *J Med Internet Res*, 27:e66279.

H. G. Wieser, W. T. Blume, D. Fish, E. Goldensohn, A. Hufnagel, D. King, M. R. Sperling, H. Lüders, and T. A. Pedley. 2001. Ilae commission report. proposal for a new classification of outcome with respect to epileptic seizures following epilepsy surgery. *Epilepsia*, 42(2):282–6.

Kevin Xie, Ryan S. Gallagher, Erin C. Conrad, Chadric O. Garrick, Steven N. Baldassano, John M. Bernabei, Peter D. Galer, Nina J. Ghosn, Adam S. Greenblatt, Tara Jennings, Alana Kornspun, Catherine V. Kulick-Soper, Jal M. Panchal, Akash R. Pattnaik, Brittany H. Scheid, Danmeng Wei, Micah Weitzman, Ramya Muthukrishnan, Joongwon Kim, and 3 others. 2022. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. *Journal of the American Medical Informatics Association*, 29(5):873–881.

Kevin Xie, Ryan S. Gallagher, Russell T. Shinohara, Sharon X. Xie, Chloe E. Hill, Erin C. Conrad, Kathryn A. Davis, Dan Roth, Brian Litt, and Colin A. Ellis. 2023. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia*, 64(7):1900–1909.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models.

Xiaoying Zhong, Siyi Li, Zhao Chen, Long Ge, Dongdong Yu, Shijia Wang, Liangzhen You, and Hongcai Shang. 2025. Considerations for patient privacy of large language models in health care: Scoping review. *J Med Internet Res*, 27:e76571.

Analyzing Prompt Design Choices in Biomedical Information Extraction for Low-Resource Languages

Ayesha Khatun and Bulut Ozler and Steven Bethard and Egoitz Laparra
College of Information Science
University of Arizona
Tucson, AZ, USA
{ayeshakhatun,kbozler,bethard,egoitz}@arizona.edu

Abstract

Despite the multilingual abilities of modern LLMs, biomedical information extraction remains challenging for low-resource, morphologically rich languages such as Bangla/Bengali and Basque. Prior investigations of prompt design and output schemas have focused on high-resource languages. To bridge this gap, we systematically analyze biomedical named entity recognition (NER) with open LLMs under multiple prompting settings on Bangla, Basque, Spanish, and English. We find that span-based extraction is more effective than BIO tagging for LLM prompting across all languages, while moving from statement-based prompting to question-based prompting has a stronger effect on the low-resource languages (e.g., +57% for Bangla and +109% for Basque) than on the high-resource languages (+28% for English and +22% for Spanish). Our breakdowns by error type show that translation-based prompting cuts Bangla hallucinations by 64% and QA-style prompting lowers Basque empty prediction errors by 61%. Our results offer practical guidance for building reliable multilingual biomedical NER systems in Bangla and Basque. We release our code at https://github.com/clulab/LRL_IE/.

1 Introduction

Research on health and medical texts is important because it affects people's lives, can reduce cost and workload, and supports better clinical decisions (Agrawal et al., 2022). However, most medical information appears in unstructured text, so medical named entity recognition (NER) is needed to identify and label key medical terms (Averly and Ning, 2025). Manual annotation, whether for direct use or for fine-tuning machine learning models, requires medical expertise, and is slow, expensive, and hard to scale. To reduce the need for manual annotation, prior work has explored

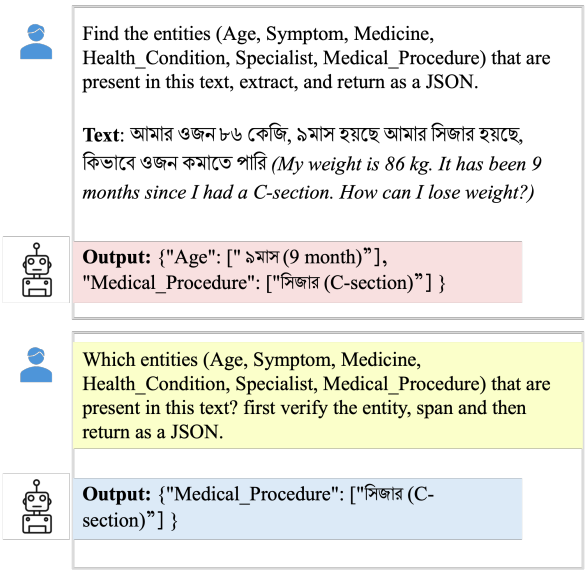


Figure 1: A statement-based prompt for biomedical NER in Bangla results in an incorrect extraction (red), while a question-based prompt results in a correct extraction (blue).

prompting and in-context learning with large language models (LLM) achieving great performance in high-resource languages like English (Ashok and Lipton, 2023; Li et al., 2023a; Li and Zhang, 2024; Wang et al., 2025; Averly and Ning, 2025). In contrast, for low-resource languages (languages with limited annotated datasets, smaller digital corpora, and fewer domain-specific resources), we still lack a systematic understanding of how different prompting strategies affect multilingual LLM performance in biomedical NER (Azime et al., 2025; Kumar et al., 2025).

We present a systematic analysis of LLM prompting strategies in Bangla/Bengali and Basque, two morphologically rich languages that pose challenging (Arnett and Bergen, 2025) low-resource settings (Bhattacharyya and Bhattacharya, 2025; López de Lacalle et al., 2020). Our analysis also includes Spanish and English as high-resource comparisons. We analyze different

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 31–44
July 3-4, 2026 ©2026 Association for Computational Linguistics

prompted output representations (beginning–inside–outside tagging (BIO; Ramshaw and Marcus, 1995) and span-based extraction), prompt variations (statement, question, explanation, and translation), and common error types (entity hallucination, over-generation, boundary mistakes, etc.). We summarize these analyses in three research questions:

- Which output format is most effective for biomedical NER? Is it consistent across Bangla, Basque, Spanish, and English?
- Which prompt variation is most effective? Is it consistent across languages?
- Which errors are most common? Which prompt choices reduce hallucination, over-generation, and boundary errors?

We find that LLMs struggle to generate output in BIO format, but can succeed with span-based extraction. We find that although the best prompting strategy is often similar across languages, with question-style generally outperforming statement-style (as in Figure 1), selecting an optimal prompt is more critical in low-resource languages, where performance degrades much more under suboptimal prompting strategies. Our analyses of these four languages offer directions for future biomedical NER research in other low-resource and multilingual settings (Pfeiffer et al., 2020; Liu et al., 2021; Choenni et al., 2023; Pham et al., 2024).

2 Related Work

Prior work has explored zero-shot and few-shot methods for Bangla across several NLP tasks (e.g., Dementieva et al. 2025; Adak et al. 2025; Li et al. 2023b; Shafayat et al. 2024); for example, Hasan et al. (2024) use 3-shot and 5-shot prompting for sentiment analysis. Prompting has also been applied to general Bangla NER: Mahtab et al. (2025) provide an English instruction prompt describing the BIO tags and rules, include 10 in-context input–output examples, and then prompt the model to label a new Bangla sentence in BIO format. Early work on Bangla biomedical/telemedicine NER (e.g., Islam et al. 2022; Sazzed 2022) primarily focused on dataset construction. More recently, Khan et al. (2023) introduce Bangla-HealthNER and evaluate fine-tuned BanglaBERT, BanglishBERT, and mBERT models, and also report substantially lower performance for zero-shot ChatGPT than for supervised fine-tuning.

For Basque, there is prior work on biomedical NER, and most of it relies on supervised training or fine-tuning with BIO-style output formats. Urbizu et al. (2022) introduce the BasqueGLUE NLU benchmark; they use standard fine-tuning and report baseline results, and for the NER task they follow the BIO annotation scheme. Zanoli et al. (2024) examine whether a multilingual clinical corpus is effective for disorder NER; they train and fine-tune supervised NER models and use IBO/BIO-style tagging. Recent work has also proposed leaderboard-style benchmarking frameworks for evaluating LLMs across multiple tasks in high-resource settings, such as Italian (Magnini et al., 2025), but comparable systematic benchmarking remains limited for Bangla and Basque biomedical NER.

However, existing work for both Bangla and Basque largely emphasizes supervised BIO tagging, and we still lack a clear picture of how prompt design and output format affect LLM-based biomedical NER in these languages. To fill this gap, we systematically evaluate multiple prompting strategies and compare BIO tagging with span-based JSON extraction.

3 Methods

Figure 2 details our NER evaluation pipeline and the different configurations we have explored in low-resource languages. We start from a base prompt that includes instructions for the task and some rules and hints to guide the LLM when annotating the entities. In all cases, the instructions are written in English (Enomoto et al., 2025) but specify the target language. Building on this base, we explore different configurations involving the output format of the annotations, whether or not to include demonstration examples, the format of these examples, or how the model should address the task.

3.1 Datasets

For our Bangla experiments, we use Bangla-HealthNER (Khan et al., 2023), a large Bengali biomedical NER dataset built from consumer health Q&A collected from a public online health platform in Bangladesh. It contains 31,783 samples and 144,136 sentences, annotated in token-level (BIO-style) format. The dataset includes seven entity types: Symptom, Health_Condition, Medicine, Specialist, Age, Dosage, and Medical_-

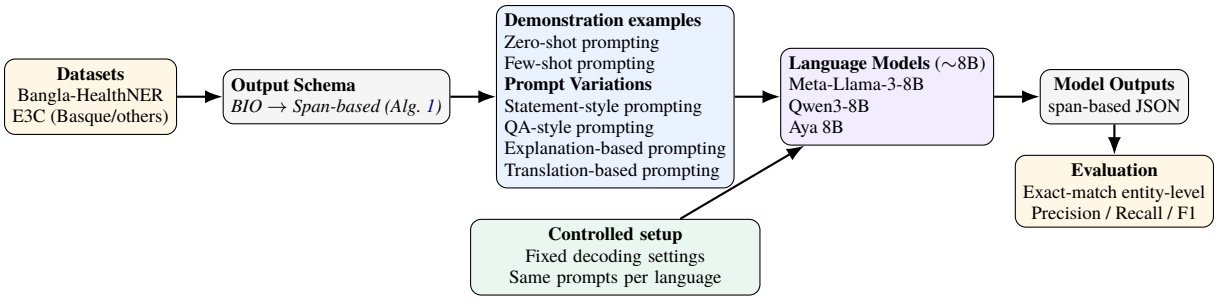


Figure 2: Overview of our biomedical NER evaluation pipeline. We vary (i) output schema (BIO vs span-level JSON) and (ii) prompting strategy, keep inference settings fixed, and evaluate with exact-match entity-level F1.

Procedure. We include Bangla-HealthNER because it is one of the largest publicly available human-annotated Bengali medical NER resources.

For Basque, Spanish, and English, we use the European Clinical Case Corpus (E3C; Magnini et al., 2020), a multilingual collection of clinical case narratives in five languages. The corpus reports clinical-case counts/tokens of English 9,533 (928K), Spanish 1,400 (531K), and Basque 122 (26K). Annotations are provided in a two-level scheme: a THYME-style layer (Styler et al., 2014) that covers time expressions events and actors, and a clinical entity layer that provides UMLS[1] Concept Unique Identifiers of different clinical disorders. From the THYME layer, we keep entities corresponding to the Patient and H-Professional (health professionals) categories. From the UMLS layer, previous work (Zanoli et al., 2024) and label all disorders with the same disorder category.

3.2 Output Schema

For all datasets, we evaluate two representations: (1) the original BIO token-level labels and (2) a span-based format (entity mentions as text spans), allowing direct comparison between tag-based and span-based extraction.

BIO format (bio) is a token-level tagging scheme for NER (Ramshaw and Marcus, 1995) that has been previously used with LLMs for Bangla NER (Mahtab et al., 2025). Given a tokenized sentence $x = (x_1, \ldots, x_n)$, dataset assigns a tag sequence $y = (y_1, \ldots, y_n)$ where each tag is one of:

$$y_i \in \{O\} \cup \{B\text{-}t,\ I\text{-}t \mid t \in \mathcal{T}\},$$

$\mathcal{T}$ is the set of entity types (e.g., AGE, SYMPTOM, etc.). Here, B marks the beginning of an entity span, I marks inside the same entity span, and O means not an entity." Figure 3 shows the prompt for the BIO tagging format.

SYSTEM_PROMPT = (
You are a biomedical NER assistant. Identify ONLY entities explicitly present in the sentence and output STRICTLY in BIO format.
Allowed output labels (BIO):
B-Age I-Age
B-Dosage I-Dosage
B-Symptom I-Symptom
B-Medicine I-Medicine
B-Health_Condition I-Health_Condition
B-Specialist I-Specialist
B-Medical_Procedure I-Medical_Procedure
O
For the text Output exactly one label per token, space-separated.
BIO_PROMPT = (
"Text: \"আমার পিত্ত থলিতে পাথর আছে। গত ২.৫ বছর যাবৎ এটা হয়েছে। যার আকার ৪ সেমি। এতদিন তীব্র কোন ব্যাথা ছিল না কিন্তু গত ৮ দিনের মধ্যে ৩ দিন পেটে তীব্র ব্যাথা হয়েছিল এবং পেটের ডানদিকে পাজরের নিচে চাপ দিলে ব্যাথা অনুভূত হয়। পিত্ত থলিতে পাথর হলে কী পেট ব্যাথার সাথে সাথে পেটের ডানদিকে পাজরের নিচে চাপ দিলে ব্যাথা অনুভূত হয়? এবং কেন এই চাপ দিলে ব্যাথা অনুভূত হয়?? অভিজ্ঞ ডাক্তারের পরামর্শ চাচ্ছি। \"\n"

Answer: O O B-Health_Condition I-Health_Condition O O
O O O O O O O
O O O O O
O O O B-Symptom O O O O O O O O O B-Symptom I-Symptom O O O O O O O O B-Symptom I-Symptom O O
O B-Specialist O O O

Figure 3: A BIO prompt for Bangla biomedical NER, where the system instruction enforces strict BIO tagging and the model outputs one label per token.

Span-based format (json) means the model does not output a tag for every token. Instead, it directly outputs the entity mentions as text spans from the sentence, usually grouped by entity type in a JSON-like structure. Many prior works adopt span-based NER models (Sohrab and Miwa, 2018; Rojas et al., 2022). In contrast, our approach predicts BIO tags and then performs BIO decoding (chunk extraction) to convert token-level predictions into span strings. Formally, BIO decoding

[1] https://www.nlm.nih.gov/research/umls/

Algorithm 1 BIO → span-level JSON

Require: Tokens $x_{1:n}$, BIO tags $y_{1:n}$, entity types $\mathcal{T}$
Ensure: $\mathcal{J}$: map $t \in \mathcal{T} \mapsto$ list of extracted span strings

$\mathcal{J}(t) \leftarrow [\,]\ \ \forall t \in \mathcal{T}$
$i \leftarrow 1$
while $i \leq n$ **do**
  if $y_i = \texttt{B-}t$ for some $t \in \mathcal{T}$ **then**
    $s \leftarrow i$
    $i \leftarrow i + 1$
    while $i \leq n$ **and** $y_i = \texttt{I-}t$ **do**
      $i \leftarrow i + 1$
    $e \leftarrow i - 1$
    $span \leftarrow \textsc{Clean}(\textsc{Detok}(x_{s:e}))$
    if $span \neq \emptyset$ **then**
      $\mathcal{J}(t) \leftarrow \mathcal{J}(t) \,\|\, [span]$
  else
    $i \leftarrow i + 1$
return $\mathcal{J}$

SYSTEM_PROMPT = (
You are a biomedical NER assistant that performs medical Named Entity Recognition .....)

Statement-style_PROMPT = (

"Find the entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) that are present in this text, extract, and return as a JSON."

"Text: \"আমার বয়স ২৪ ওজন ৫৮ কেজি আমার কিছুখন পর পর প্রসাব হয় খুবকম এখন আমি কি করতে পারি। \"\n"
"Answer: {\"Age\":[], "
"\"Symptom\":[\"প্রসাবের রাস্তায় জ্বালাপোড়া\",\"ব্যথা\",\"প্রসাবে মাঝে মাঝে গন্ধ\","
"\"Medicine\":[], "
"\"Health_Condition\":[], "
"\"Specialist\":[], "
"\"Medical_Procedure\":[]}\n\n")

Figure 4: A statement-style prompt for Bangla biomedical NER, where the yellow-highlighted text tells the model to find predefined medical entity types in the input text.

yields a set of extracted spans:

$$\hat{\mathcal{S}} = \left\{ (t, x_{s:e}) \;\middle|\; \begin{array}{l} y_s = \texttt{B-}t, \forall i \in (s{+}1, \ldots, e) : y_i = \texttt{I-}t \\ y_{e+1} \neq \texttt{I-}t \end{array} \right\}$$

where $\mathcal{T}$ is the set of entity types and $x_{s:e}$ is the token span reconstructed into text. We used Algorithm 1 to detokenize each extracted span and append it to $\mathcal{J}(t)$. Code is available at https://anonymous.4open.science/r/LRL_IE-1959/.

3.3 Demonstration Examples

Zero-shot prompting (zero) is a strategy where we provide only the task instruction and the list of entity types, without any labeled examples (Liu et al., 2023). We provided detailed guidelines to the model, e.g., instructing it to skip negated symptoms, not confuse duration with age, not treat lab/test names as symptoms, and to select short, clean spans with no duplicates. This setting tests whether the model can perform biomedical extraction with only instructions and no demonstrations.

Few-shot prompting (few) is a way to use an LLM without training it, where we show the model a small number of labeled examples (Pan et al., 2023; Cheng et al., 2025) inside the prompt and then ask it to do the same task for a new input. Such labeled examples are known as the *support set*, $\mathcal{S}_K$, where K labeled examples are included in the prompt. Formally, we combine the base prompt P_{base} with this support set, and find the LLM's most probable output, $\hat{o}_{\text{FS}}$, given this combined prompt:

$$P_{\text{FS}} = P_{\text{base}} \oplus \mathcal{S}_K$$
$$\hat{o}_{\text{FS}} = \arg\max_o \; p(o \mid x, P_{\text{FS}})$$

The few-shot prompt contains $K = 9$ in-context examples in the target language. Since there are nine entity types in total, we set $K = 9$ and designed the examples so that each entity type appears at least once. The selected set includes both short and long examples.

3.4 Prompt Variations

Statement-style prompting (stmt) appends an statement of the task s to the prompt while using the same support set $\mathcal{S}_K$:

$$P_{\text{QA}} = P_{\text{base}} \oplus \mathcal{S}_K \oplus s$$
$$\hat{o}_{\text{QA}} = \arg\max_o \; p(o \mid x, P_{\text{QA}})$$

This approach presents the biomedical entity extraction task as a statement, such as "Find the entities (Age, Symptom, …) that are present in this text.", as shown in Figure 4.

Question-style prompting (qa) appends an explicit question q to the prompt while using the same support set $\mathcal{S}_K$:

$$P_{\text{QA}} = P_{\text{base}} \oplus \mathcal{S}_K \oplus q$$
$$\hat{o}_{\text{QA}} = \arg\max_o \; p(o \mid x, P_{\text{QA}})$$

This approach presents the biomedical entity extraction task as a direct question, such as "Which entities of types (Age, Symptom, …) appear in this text?", as shown in Figure 5.

Explanation-based prompting (expl) is a prompting strategy (Figure 6) that provides a brief description of what entities are likely present in the current text. First, the prompt explains the label boundaries (e.g., Health_Condition is a diagnosis or stated condition, while Symptom is a complaint). Then, for each example, we add a short natural-language description in English that highlights the relevant cues in the input. We have

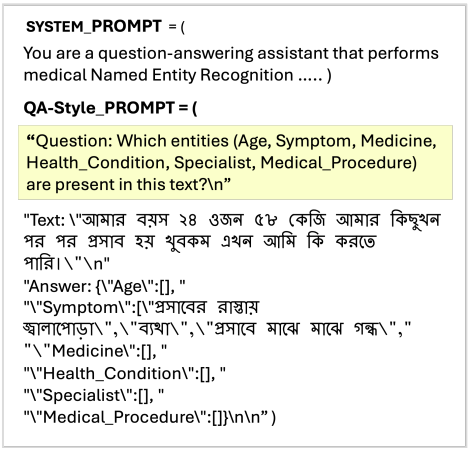


Figure 5: A question-style prompt for Bangla biomedical NER, where the yellow-highlighted text explicitly asks the model to identify predefined medical entity types in the input text.

used explanation only for examples, not the entire prompt. This prompt uses the same support set $\mathcal{S}_K$ and adds label definitions $D = \{d_\ell\}_{\ell \in \mathcal{L}}$:

$$P_{\text{Desc}} = P_{\text{base}} \oplus \mathcal{S}_K \oplus D$$
$$\hat{o}_{\text{Desc}} = \arg\max_o \; p(o \mid x, P_{\text{Desc}})$$

Translation-based prompting (trans) is a process of prepending an English rendering (Figure 7) of the input while also keeping the original Bangla/Basque text in the prompt. This encourages the model to reason in English, which can help when the model is stronger in English than in the target language. However, this approach can fail when translation paraphrases the meaning, drops details, mistranslates medical terms, or changes span boundaries. Therefore, we treat translation-based prompting as a practical baseline rather than a guaranteed improvement. Let $T(\cdot)$ be a translation function (e.g., Bangla/Basque $\rightarrow$ English). We use the same support set $\mathcal{S}_K$, but the input includes the translated text:

$$x' = T(x)$$
$$P_{\text{Trans}} = P_{\text{base}} \oplus \mathcal{S}_K$$
$$\hat{o}_{\text{Trans}} = \arg\max_o \; p(o \mid x', P_{\text{Trans}})$$

3.5 Language Models

We primarily conduct our experiments using **Meta-Llama-3-8B** (Grattafiori et al., 2024), a widely adopted open-weights model with strong general performance. However, to test the effect of different model families, we also evaluate the best-performing prompt strategy under LLama-3 on two additional models: **Qwen3-8B** (Yang et al., 2025) and **Aya** (Aryabumi et al., 2024). We include Qwen3 because it is a competitive recent model that often performs particularly well on high-resource languages such as English, and Aya because it is designed for multilingual use and thus provides a useful contrast for low-resource settings. To ensure a fair comparison, we select similar size models (all in the $\sim$8B parameter range) and and keep inference settings fixed.

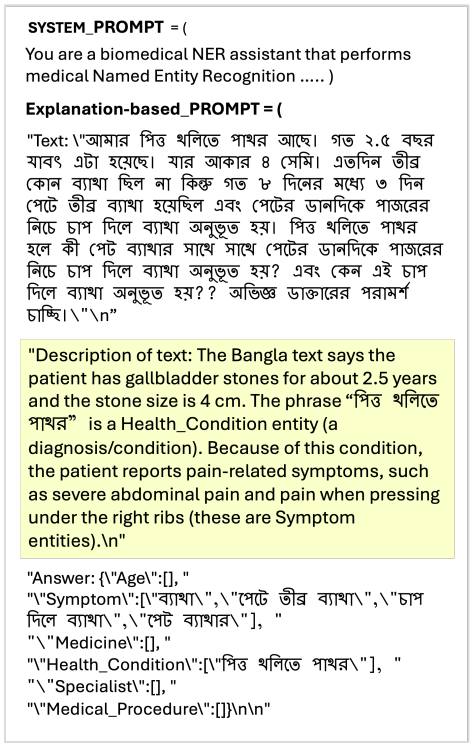


Figure 6: An explanation-based prompt for Bangla biomedical NER, where the yellow-highlighted text provides a natural-language explanation of the clinical context and entity semantics.

4 Experimental Results

Our analyses are presented below. As datasets for different languages are by necessity different, we focus on relative gains, not absolute values, when looking across languages.

4.1 Output Format Analysis

Table 1 shows a large gap between BIO tagging and span-based extraction, indicating that the output schema strongly affects LLM-based NER. Across languages, switching from *bio_few* to *json_few_stmt* yields large F1 gains: +538% (Bangla; 0.054 $\rightarrow$ 0.345), +200% (Basque; 0.104 $\rightarrow$ 0.312), +2242% (Spanish; 0.019 $\rightarrow$ 0.445), and +339% (English; 0.120 $\rightarrow$ 0.527). BIO tagging is likely harder for LLMs because the model must label

SYSTEM_PROMPT = (

You are a biomedical NER assistant that performs medical Named Entity Recognition .....)

Translation-based_PROMPT = (

"Text: \"আমার পিত্ত থলিতে পাথর আছে। গত ২.৫ বছর যাবৎ এটা হয়েছে। যার আকার ৪ সেমি। এতদিন তীব্র কোন ব্যাথা ছিল না কিন্তু গত ৮ দিনের মধ্যে ৩ দিন পেটে তীব্র ব্যাথা হয়েছিল এবং পেটের ডানদিকে পাজরের নিচে চাপ দিলে ব্যাথা অনুভূত হয়। পিত্ত থলিতে পাথর হলে কী পেট ব্যাথার সাথে সাথে পেটের ডানদিকে পাজরের নিচে চাপ দিলে ব্যাথা অনুভূত হয়? এবং কেন এই চাপ দিলে ব্যাথা অনুভূত হয়?? অভিজ্ঞ ডাক্তারের পরামর্শ চাচ্ছি। \"\n"

"Translation of the text: I have gallstones in my gallbladder. This has been happening for the last 2.5 years. The stone size is 4 cm. I did not have any severe pain for a long time, but within the last 8 days, I had severe stomach pain on 3 days, and I feel pain when I press under the right rib on the right side of my abdomen. If there are gallstones, does abdominal pain occur along with pain when pressing under the right rib on the right side? And why do I feel pain when I press there?? I want advice from an experienced doctor.\n"

"Answer: {\"Age\":[], "
"\"Symptom\":[\"ব্যাথা\",\"পেটে তীব্র ব্যাথা\",\"চাপ দিলে ব্যাথা\",\"পেট ব্যাথার\"], "
"\"Medicine\":[], "
"\"Health_Condition\":[\"পিত্ত থলিতে পাথর\"], "
"\"Specialist\":[], "
"\"Medical_Procedure\":[]}\n\n"

Figure 7: A translation-based prompt for Bangla biomedical NER, where the yellow-highlighted text shows the English translation of the original Bangla clinical text used for entity extraction.

Prompt	Bangla	Basque	Spanish	English
bio_few	0.054	0.104	0.019	0.120
json_few	**0.345**	**0.312**	**0.445**	**0.527**

Table 1: Biomedical NER Performance across output schemas with Llama3-8B. In both cases, we apply few-shot prompting.

every token in the exact order, with even one missing/extra token breaking the whole alignment. Span-based extraction is easier because the model can simply copy the entity text spans and return them as clean JSON, which matches how LLMs naturally answer.

4.2 Prompt Variation Analysis

Table 2 shows that QA-style prompting performs best across all languages. It even outperforms detailed label explanations (*expl*) and translation of the input to English (*trans*), despite using only simple WH-questions. The strong performance of QA-style prompting is likely because modern LLMs are instruction tuned on many question-answer pairs. Prior work reformulating NER as machine reading comprehension similarly finds that QA-style formulations improve NER, attributing gains

Prompt	Bangla	Basque	Spanish	English
json_zero	0.291	0.241	0.403	0.422
json_few_stmt	0.345	0.312	0.496	0.527
json_few_qa	**0.458**	**0.503**	**0.528**	**0.541**
json_few_trans	0.347	0.441	0.494	0.313
json_few_expl	0.417	0.386	0.477	0.304

Table 2: Prompting Strategy Comparison for Multilingual Biomedical NER Using Meta-Llama-3-8B

to query conditioning and (when available) semantically informative queries that encode entity-type knowledge (Li et al., 2020).

Prompting strategy has a stronger effect on low-resource languages than on high-resource languages: moving from zero-shot to question-style prompting improves Bangla by 57.4% and Basque by 108.7%, while Spanish improves by 21.8% and English by only 28.2%.

4.3 Entity Type Analysis

Table 3 and Table 4 break down performance of prompting strategies by named entity types for Bangla and Basque, respectively. In both languages, question-style prompting is best for most types, with the largest absolute gains for MEDICAL_PROCEDURE in Bangla (37.8%) and H-PROFESSIONAL in Basque (37.4%), indicating that such prompting is especially helpful for rarer or harder-to-extract categories.

Occasionally, explanation or translation-based prompting outperforms question-based prompting. Explanation-based prompting helps Bangla SYMPTOMS and translation-based prompting helps Bangla SPECIALISTS. But the gains are modest over question-based prompting, and translation-based prompting fails badly for the AGE category.

4.4 Analysis Under the Best Prompt

Table 5 compares different LLMs under the best-performing (question-style) prompt. Overall, Llama-3-8B performs best on Bangla (F1 = 0.494) and Spanish (F1 = 0.529), while Qwen3-8B achieves the highest score on English (F1 = 0.541) and Basque (F1 = 0.550). Aya performs worst on all languages. This suggests that Qwen may have stronger English and Basque knowledge from pretraining, while Llama may have stronger Spanish and Bangla knowledge, and both Qwen and Llama likely have better multilingual and/or biomedical domain knowledge than Aya. These results reveal the importance of LLM selection when working with low-resource languages.

	Age	Symptom	Medicine	Health_Condition	Specialist	Medical_Procedure
json_zero	0.421	0.067	0.400	0.400	0.514	0.133
json_few_stmt	0.200	0.198	0.490	0.400	0.621	0.167
json_few_qa	**0.471**	0.215	**0.600**	**0.533**	0.600	**0.546**
json_few_trans	0.010	0.196	0.560	0.333	**0.625**	0.222
json_few_expl	0.235	**0.253**	**0.600**	0.519	0.500	0.364

Table 3: Entity-type–wise F1 scores for Bangla biomedical NER under five prompting strategies. Bold indicates the best F1 per entity type.

	Disorder	Patient	H-Professional
json_zero	0.215	0.368	0.140
json_few_stmt	0.314	0.441	0.183
json_few_qa	**0.427**	**0.515**	**0.558**
json_few_trans	0.380	0.410	0.320
json_few_expl	0.411	0.409	0.328

Table 4: Entity-type-wise F1 scores for Basque biomedical NER under five prompting strategies. Bold indicates the best F1 per entity type.

Model	Bangla	Basque	Spanish	English
Llama-3-8B	**0.494**	0.503	**0.529**	0.457
Qwen3-8B	0.399	**0.550**	0.514	**0.541**
Aya	0.229	0.282	0.349	0.318

Table 5: Prompting Strategy Comparison for Multilingual Biomedical NER Using Meta-Llama-3-8B and question-style prompting (json_few_qa).

5 Error Analysis

We inspected the errors of the best performing models in Bangla and Basque. Below are the main failure modes, with the operational definitions used in our evaluation. Examples of errors are given in Appendix D and Appendix E.

Hallucination The gold annotation contains no entities, but the model outputs one or more spans.

All-Missed The gold contains entities, but the model returns an empty JSON. This is the most severe recall failure.

Missed Entities The model extracts some entities but misses others, without adding extra entities.

Extra Entities The model predicts additional entities that are not in the gold but does not miss gold entities.

Type Confusion The predicted entity span matches the gold text, but the assigned type is wrong.

Mixed Errors The model both adds and misses entities. Extra-dominant cases indicate over-extraction, while missed-dominant cases indicate under-extraction.

Boundary Mismatch The model captures the right concept but with different boundaries, becoming both a false positive and a false negative under exact-match scoring.

Figure 8 (top) shows clear differences across prompting methods for Bangla biomedical NER errors. Although we have seen that QA-style prompts perform better on average, here we see that translation-based prompting reduces over-generation: it has the lowest HALLUCINATION (2.6%) and comparatively low BOUNDARY MISMATCH (13%). For span-level precision, question-style prompting yields the largest BOUNDARY MISMATCH (29%), suggesting it identifies the right concept but struggles to copy the exact multiword span. Thus, question-style prompting helps the model extract more entities, while translation-based prompting helps reduce hallucination, extra entities, and boundary mismatch errors.

For Basque (Figure 8, bottom), across all prompts, HALLUCINATION dominates (38–52% of errors), indicating that over-generation is the primary failure mode in Basque. However, the prompts shift the secondary errors: statement-based prompting (*json_few_stmt*) reduces HALLUCINATION (37.5%) compared to the other methods, but it increases ALL-MISSED (14.8%), suggesting a more conservative behavior that sometimes fails to extract anything. In contrast, question-style prompting extracts more aggressively and keeps ALL-MISSED low (5.8%), but this comes with higher EXTRA ENTITIES (8.6%) and substantial BOUNDARY MISMATCH (18.5%).

6 Conclusion and Future Work

We systematically analyzed prompt design choices and output schemas for multilingual biomedical NER, focusing on the low-resource, morphologically rich languages Bangla and Basque while comparing them against the high-resource languages Spanish and English. We find that question-style prompting is the most consistent way to improve

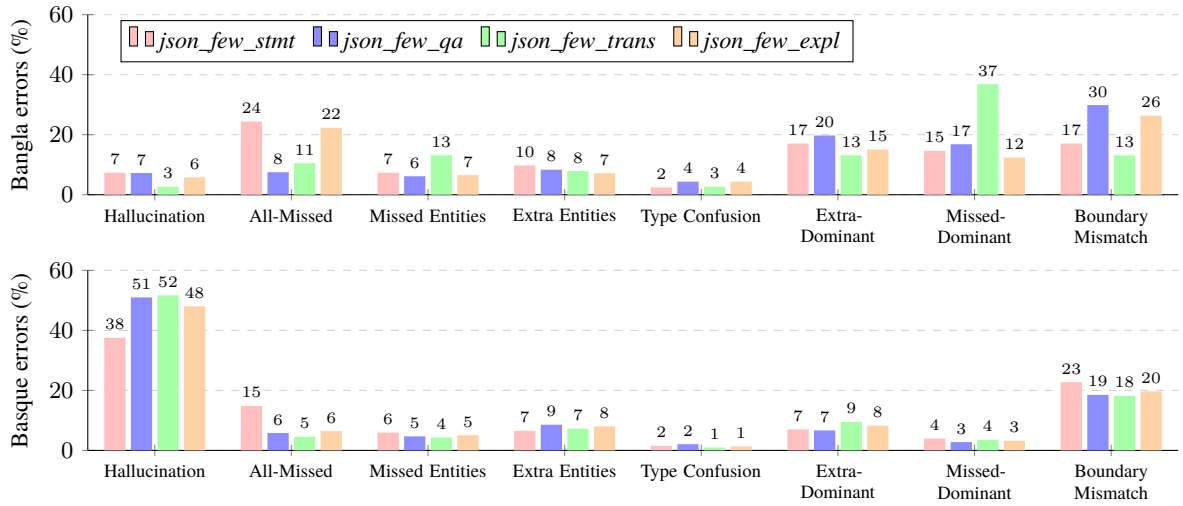


Figure 8: Error type distribution for LLM-extracted entities for Bangla (top) and Basque (bottom).

exact-match entity-level F1 across all languages. We also find that selecting an optimal prompt is more critical in low-resource languages, where performance degrades much more under suboptimal prompting strategies. Our error analysis shows that question-style prompts result in fewer empty outputs, while translation-based prompting can reduce hallucinations and boundary errors in some settings.

We also find that BIO-tagging output is less suitable for LLMs than text-span based output. This contrasts with available datasets: not only biomedical NER, but nearly all Bangla NER datasets are annotated in the BIO format (e.g., Islam et al., 2022; Sazzed, 2022; Khan et al., 2023; Mahtab et al., 2025). BIO format is fragile: the model must keep the exact token order and token count, and even one extra or missing token breaks alignment. This becomes worse when tokenization is difficult, as in morphologically rich languages like Bangla and Basque, where suffixes and punctuation frequently change token boundaries. In contrast, text-span-based extraction asks the model to list the entity strings in the sentence, which matches how LLMs naturally respond and is easier to parse and store as JSON.

In the future, we plan to test richer question prompts that include short label definitions and examples (not only WH-questions) to see whether they further reduce boundary mismatches and type confusion. We also want to study better post-processing and evaluation for near-miss spans and explore lightweight adaptation methods that choose the best prompt per language and entity type. Finally, expanding to more low-resource languages and more medical datasets will help confirm how general these findings are.

Limitations

A key limitation of this study is that we evaluate only three open LLMs around 8B parameters and two datasets, Bangla HealthNER and Basque E3C, so further work is needed to determine whether the findings generalize to other model sizes, closed models, or biomedical text styles. We also do not cover additional low-resource languages because validating prompts, examples, and error cases reliably requires language expert support. In addition, we did not evaluate paid or proprietary GPT models, and our strict exact match scoring can penalize near-correct span boundaries in morphologically rich languages. Finally, since different datasets are used for each language, cross-language comparisons should be interpreted cautiously, as proportional improvements may still reflect dataset-specific characteristics rather than true language differences.

References

Sayantan Adak, Pauras Mangesh Meher, Paramita Das, and Animesh Mukherjee. 2025. REVerSum: A multi-staged retrieval-augmented generation method to enhance Wikipedia tail biographies through personal narratives. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 732–750, Abu Dhabi, UAE. Association for Computational Linguistics.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language

models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Dhananjay Ashok and Zachary C. Lipton. 2023. PromptNER: Prompting For Named Entity Recognition. *arXiv preprint*. ArXiv:2305.15444 [cs].

Reza Averly and Xia Ning. 2025. Entity decomposition with filtering: A zero-shot clinical named entity recognition framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2935–2951, Albuquerque, New Mexico. Association for Computational Linguistics.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266, Albuquerque, New Mexico. Association for Computational Linguistics.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. BanglaByT5: Byte-level modelling for Bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5551–5560, Suzhou, China. Association for Computational Linguistics.

Xiang Cheng, Chengyan Pan, Minjun Zhao, Deyang Li, Fangchao Liu, Xinyu Zhang, Xiao Zhang, and Yong Liu. 2025. Revisiting chain-of-thought prompting: Zero-shot can be stronger than few-shot. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13533–13554, Suzhou, China. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, 49(3):613–641.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818, Torino, Italia. ELRA and ICCL.

Tanvir Islam, Sakila Mahbin Zinat, Shamima Sukhi, Zakir Hossain Zamil, Aynur Nahar, and M. F. Mridha. 2022. An attention-based medical NER in the Bengali language. In G. Mathur and 1 others, editors, *Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications*, Algorithms for Intelligent Systems, pages 131–140. Springer Nature Singapore.

Alvi Khan, Fida Kamal, Nuzhat Nower, Tasnim Ahmed, Sabbir Ahmed, and Tareque Chowdhury. 2023. NERvous about my health: Constructing a Bengali medical named entity recognition dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5768–5774, Singapore. Association for Computational Linguistics.

Somnath Kumar, Vaibhav Balloli, Mercy Ranjit, Kabir Ahuja, Sunayana Sitaram, Kalika Bali, Tanuja Ganu,

and Akshay Nambi. 2025. Bridging the language gap: Dynamic learning strategies for improving multilingual performance in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9209–9223, Abu Dhabi, UAE. Association for Computational Linguistics.

Mingchen Li, Yang Ye, Jeremy Yeung, Huixue Zhou, Huaiyuan Chu, and Rui Zhang. 2023a. W-procer: Weighted Prototypical Contrastive Learning for Medical Few-Shot Named Entity Recognition. *arXiv preprint*. ArXiv:2305.18624 [cs].

Mingchen Li and Rui Zhang. 2024. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023b. Crosslingual retrieval augmented in-context learning for Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 136–151, Singapore. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Guizhen Chen, Xiaobao Wu, Anh Tuan Luu, Chip Hong Chang, and Lidong Bing. 2023. Zero-shot text classification via self-supervised tuning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1743–1761, Toronto, Canada. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Maddalen López de Lacalle, Xabier Saralegi, and Iñaki San Vicente. 2020. Building a task-oriented dialog system for languages with no training data: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2796–2802, Marseille, France. European Language Resources Association.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The E3C project:collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 190–196, Bologna, Italy. CEUR Workshop Proceedings.

Bernardo Magnini, Marco Madeddu, Michele Resta, Roberto Zanoli, Martin Cimmino, Paolo Albano, and Viviana Patti. 2025. A leaderboard for benchmarking LLMs on Italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 636–646, Cagliari, Italy. CEUR Workshop Proceedings.

Md. Motahar Mahtab, Faisal Ahamed Khan, Md. Ekramul Islam, Md. Shahad Mahmud Chowdhury, Labib Imam Chowdhury, Sadia Afrin, Hazrat Ali, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2025. BanNERD: A benchmark dataset and context-driven approach for Bangla named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6807–6828, Albuquerque, New Mexico. Association for Computational Linguistics.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Matias Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. 2022. Simple yet powerful: An overlooked architecture for nested named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2108–2117, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Salim Sazzed. 2022. BanglaBioMed: A biomedical named-entity annotated corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Sheikh Shafayat, H M Quamran Hasan, Minhajur Rahman Chowdhury Mahim, Rifki Afina Putri, James Thorne, and Alice Oh. 2024. BEnQA: A question answering benchmark for Bengali and English. In *Findings of the Association for Computational Linguis-*

tics: ACL 2024, pages 1158–1177, Bangkok, Thailand. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

IV Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. BasqueGLUE: A natural language understanding benchmark for Basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Roberto Zanoli, Alberto Lavelli, Daniele Verdi do Amarante, and Daniele Toti. 2024. Assessment of the E3C corpus for the recognition of disorders in clinical texts. *Natural Language Engineering*, 30(4):851–869.

A BIO Tagging vs Span-Based

The left panel of Figure 9 shows span-based extraction, where entities are directly returned as text spans grouped by type in a JSON structure, while the right panel shows token-level BIO tagging, which requires assigning a label to every token and maintaining strict alignment. This example illustrates how span-based formats provide a simpler and more robust representation for LLM-based NER, especially for morphologically rich languages like Bangla.

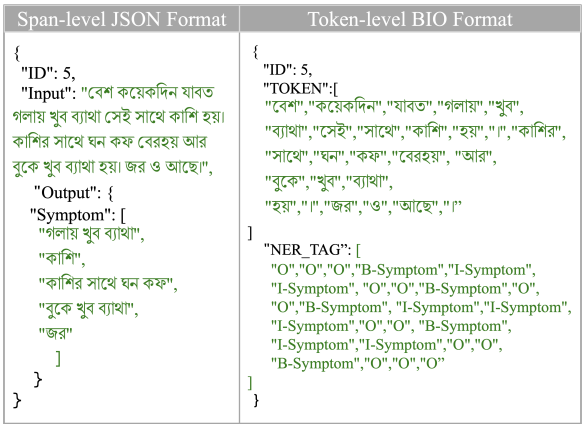


Figure 9: Comparison of span-level JSON and token-level BIO formats for Bangla biomedical NER.

B Prompting Behavior and Qualitative Analysis for Bangla

Table 6 illustrates qualitative differences across three prompting strategies for Bangla biomedical NER. In the question-style prompt, the model correctly identifies most entities but misses the HEALTH_CONDITION (ইনফেকশন), showing that QA prompting can still suffer from recall gaps. In contrast, translation-based prompting leads to overgeneration: the model incorrectly treats common home remedies such as গরম পানি (hot water) and মধু as MEDICINE, likely due to reasoning in English and broader semantic interpretation. Explanation-based prompting improves recall but introduces type confusion, as the model incorrectly labels the polite address স্যার (“Sir”) as a SPECIALIST. These examples highlight that while extended prompting strategies can help extraction, they also introduce distinct error patterns depending on how semantic cues are framed.

C Prompting Behavior and Qualitative Analysis for Basque

Table 7 presents qualitative differences in model behavior across three prompting strategies for Basque biomedical NER. Under the question-style prompt, the model successfully identifies major disorders but exhibits boundary and normalization issues, such as over-specifying severity or partially mismatching gold entity forms. Translation-based prompting generally improves coverage by leveraging English semantics, but it can introduce label drift, as closely related clinical concepts are normalized differently from the gold labels. Explanation-based prompting further increases re-

E.g.	Prompt Type	Prompt	Llama3-8B	Gold Label
1	json_few_-qa	You are a question-answering assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. Question: Which entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in this text? **Example 1:** এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না। **Answer:** Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির. **Now extract all entities for this text:** ধন্যবাদ আপনার প্রশ্নের জন্য। ফাটিয়ে দিবেন না। ইনফেকশন হয়ে যাবে। আপাতত এলারজি উদ্রেককারী খাবার, ধূলাবালি এড়িয়ে চলুন। এলাট্রল খান। না দেখে সমাধান দেয়া যাচ্ছে না। সরাসরি চর্মরোগ বিশেষজ্ঞ দেখিয়ে পরামর্শ নিতে হবে ধন্যবাদ	Medicine: এলাট্রল Specialist: চর্মরোগ বিশেষজ্ঞ	Medicine: এলাট্রল Health_Condition: ইনফেকশন Specialist: চর্মরোগ বিশেষজ্ঞ
2	json_few_-trans	You are a NER assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. **Example 1:** এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না। English translation of the Example: Because of allergy problems, when I go out in the sun my body tingles, and I feel a crawling sensation inside my head. In the morning after waking up I sneeze continuously. I have never taken any allergy medicine regularly. In this case, what can I do? Because of this allergy, I cannot do many daily tasks. Now findout the entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in this text. **Answer:** Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির. **Now extract all entities for this text:** আপনার বাচ্চাকে আদা, মধু, গরম পানি খাওয়ান। সিরাপ এমব্রোক্স খাওয়াতে পারেন আধা চামচ করে তিন বার একজন শিশু বিশেষজ্ঞ কে দেখিয়ে নিন।	Medicine: আদা | গরম পানি | মধু | সিরাপ এমব্রোক্স ; Specialist: শিশু বিশেষজ্ঞ	Medicine: সিরাপ এমব্রোক্স ; Specialist: শিশু বিশেষজ্ঞ
3	json_few_-expl	You are a NER assistant that performs medical Named Entity Recognition (NER). For each question, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. **Example 1:** এলার্জির সমস্যার কারণে রোদে গেলে গা চিটমিট করে, মাথার ভিতরে কিলবিল করে। সকালে ঘুম থেকে উঠলে অনবরত হাঁচি হয়। কখনো নিয়মিত কোনো এলার্জির ওষুধ খাইনি। এক্ষেত্রে আমি কি করতে পারি? এনার্জির কারণে অনেক দৈনন্দিন কাজ করতে পারি না। Explanation of this Example: The Bangla text clearly describes an allergy problem, which is a Health_Condition entity (“এলার্জির”). It also describes allergy-related symptoms, such as skin discomfort in the sun, crawling sensation in the head, and continuous sneezing (Symptom). It mentions “এলার্জির ওষুধ” only in a generic way, but it still refers to medicine use (Medicine). No age, specialist, or procedure is stated. Now read the Bangla text and check the Answer, then identify which entities (Age, Symptom, Medicine, Health_Condition, Specialist, Medical_Procedure) are present in the text. **Answer:** Symptom – রোদে গেলে গা চিটমিট করে | মাথার ভিতরে কিলবিল করে | হাঁচি; Medicine – এলার্জির ওষুধ. Health_Condition – এলার্জির. **Example 2:** রোগীর বয়স ৫৫ বছর। তিনি ডায়াবেটিস ও উচ্চ রক্তচাপের রোগী এবং মেটফরমিন ও লোসারটান খাচ্ছেন। **Answer:** Age – ৫৫ বছর. Health_Condition – ডায়াবেটিস | উচ্চ রক্তচাপ. **Now extract all entities for this text:** আমার বয়স ১৩ আমার পচন্ড পায়ের হার এর বেথা। আমি কি ঔষধ খাব বলেন স্যার। Calcium BP 500mg কিনছি খাব কি স্যার।	Age: ১৩ বছর ; Symptom: পচন্ড পায়ের হার এর বেথা ; Medicine: Calcium BP 500mg ; Specialist: স্যার	Age: ১৩ ; Symptom: পায়ের হার এর বেথা ; Medicine: Calcium BP 500mg

Table 6: Example prompt templates for Bangla biomedical NER. We show three prompt variants (question-style, translation-based, and explanation-based) with one in-context demonstration and one target Bangla input. For each example, we report the Meta-Llama-3-8B output and the corresponding gold labels for comparison.

E.g.	Prompt Type	Prompt	Llama3-8B	Gold Label
1	json_few_-qa	You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. Question: Which entities (Disorder, Patient, H-Professional) are present in this text? **Example 1:** Text: *"-Eskuineko besapea:aurkikuntza patologikorik ez."* **Answer:** {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []} **Now extract all entities for this text:** *Estenosi aortiko larriari bihotz-gutxiegitasuna eta sinkopea gehitu zaizkionez, ebakuntza egiteko baldintzak betetzen ditu, eta Bihotz-hodietako Kirurgiako Zerbitzuarekin harremanetan jarri gara.*	Disorder: bihotz-gutxiegitasuna | estenosi aortiko larria | sinkopea ; H-Professional: Bihotz-hodietako Kirurgiako Zerbitzuarekin	Disorder: Estenosi aortiko | bihotz-gutxiegitasuna | sinkopea]
2	json_few_-trans	You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. **Example 1:** Text: *"-Eskuineko besapea:aurkikuntza patologikorik ez."* English translation of the text: "-Right armpit: no pathological findings." **Answer:** {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []} **Now extract all entities for this text:** *Txankro sifilitikoa/Sifilis primarioa.*	Disorder: Sifilis primarioa | Txankro sifilitikoa	Disorder: Sifilis | Txankro sifilitikoa
3	json_few_-expl	You are an assistant that performs medical Named Entity Recognition (NER) in Basque. For each input text, identify ONLY the entities that are explicitly present in the provided text and answer STRICTLY in JSON. **Example 1:** Text: *"-Eskuineko besapea:aurkikuntza patologikorik ez."* Description of text: A clinical examination of a right armpit with no health disorders or pathological findings identified. You have to find out which entities (Disorder, Patient, H-Professional) are present in the text. **Answer:** {'Disorder': ['aurkikuntza'], 'Patient': [], 'H-Professional': []} **Now extract all entities for this text:** *Edemarik ez; ez dago bena sakonetako tronbosi-zeinurik.*	Disorder: tronbosi-zeinurik	Disorder: Edemarik | bena sakonetako tronbosi-zeinurik

Table 7: Example prompt templates for Basque biomedical NER. We show three prompt variants (question-style, translation-based, and explanation-based) with one in-context demonstration and one target Basque input. For each example, we report the Meta-Llama-3-8B output and the corresponding gold labels for comparison.

Error Type	Input Example	Gold Output	Pred Output
Hallucinations	ধন্যবাদ আপনাকে প্রশ্নের জন্য। জ্বর ৯৯ হলে ওষুধ খাওয়াবেন। নয়ত দরকার নেই। ধন্যবাদ	*Empty*	Symptom: জ্বর
All-Missed	আমি। আমার বয়স 27 বছর। সাম্প্রতিক সময়ে আমি আমার শরীরে অস্বস্...ান্ত হয়ে গেছি, আগের থেকে শরীরের ওজন বৃদ্ধি পেয়েছে, পেটে প্র...	Age: 27 বছর ; Symptom: একাধারে মাটিতে বসে থাকলে মাজার ...| শরীরের ওজন বৃদ্ধি | হাটলেই মনে হয় ক্লান্ত হয়ে	*Empty*
Missed Entities	total cholesterol 200 mg / dl serum triglycerides 553 mg / dl…how to lower triglycerides? am i prone to heart disease?	Symptom: HDL cholesterol 25 mg | LDL cholesterol 93 mg | cholesterol…| serum triglycerides 553 mg ; Health_Condition: heart disease	Health_Condition: heart disease
Extra Entities	আপনাকে ধন্যবাদ প্রশ্ন করার জন্য। আপনি গরম পানির ধোয়া নাকে নিয়...রাতে একটা করে খান। একজন মেডিসিন বিশেষজ্ঞ পরামর্শ নিন। ধন্যবাদ	Medicine: এলাট্রল ; Specialist: মেডিসিন বিশেষজ্ঞ	Medicine: এলাট্রল | গরম পানির ধোয়া ; Specialist: মেডিসিন বিশেষজ্ঞ
Type Confusion	আমার স্ত্রী বয়স ২৪ ও ১৪ দিনের প্রেগন্যান্ট অবস্থা MM Kit খায় ...আল্টাস্রোফ্রাগি করাতে চাই, দয়া করে আমাকে পরিক্ষার নাম টা লিখে দিবেন	Age: ২৪ ; Symptom: bleeding আনে বেশি যাচ্ছে ; Medicine: MM Kit ...; Health_Condition: মাসিক | ১৪ দিনের প্রেগন্যান্ট	Age: ১৪ দিন | ২৪ ; Symptom: bleeding ; Medicine: MM Kit ; Health_Condition: প্রেগন্যান্ট অবস্থা ; Medical_Procedure: আল্টাস্রোফ্রাগি
Mixed Errors	Thank you for your question. Your serum Triglyceride is sligh…with your reports. Avoid fatty food, carbohydrate. Thank you.	Symptom: serum Triglyceride is slightly raised	Medicine: drug ; Specialist: doctor
Boundary Mismatch	বেশ কয়েকদিন যাবত গলায় খুব ব্যাথা সেই সাথে কাশি হয়। কাশির সাথে ঘন কফ বেরহয় আর বুকে খুব ব্যাথা হয়। জর ও আছে।	Symptom: কাশি | কাশির সাথে ঘন কফ | গলায় খুব ব্যাথা | জর | বুকে খুব ব্যাথা	Symptom: কাশি | খুব ব্যাথা | ঘন কফ | জর | বুকে খুব ব্যাথা

Table 8: Representative QA-prompt error examples for Bangla biomedical NER (Llama3.1-8B). Each row shows one real instance of a major error category under exact-match span scoring.

Error Type	Input Example	Gold Output	Pred Output
Hallucinations	*Burua eta lepoa.*	*Empty*	Disorder: Burua eta lepoa
All-Missed	*Ez dago aurkikuntza patologikorik.*	Disorder: aurkikuntza	*Empty*
Missed Entities	*Estenosi aortiko larria eta bihotz-gutxiegitasuna.*	Disorder: Estenosi aortiko larria | bihotz-gutxiegitasuna	Disorder: bihotz-gutxiegitasuna
Extra Entities	*Sabela.*	*Empty*	Patient: Sabela
Type Confusion	*Biguna eta zanpagarria.*	*Empty*	Disorder: Biguna | Disorder: zanpagarria
Mixed Errors	*Biriketako murmurio normala.*	*Empty*	Disorder: Biriketako murmurio
Boundary Mismatch	*Ez dago aurkikuntza patologikorik.*	Disorder: aurkikuntza patologikorik	Disorder: aurkikuntza

Table 9: Representative QA-prompt error examples for Basque biomedical NER (Meta-Llama-3-8B). Each row illustrates a distinct error category under exact-match span evaluation.

call by encouraging semantic inference, yet it also leads to over-generation, with the model extracting implicit or negated findings (e.g., absence of edema or thrombosis) as entities. Overall, these examples demonstrate that richer prompting strategies can enhance extraction in Basque but also introduce systematic errors related to semantic inference, normalization, and negation handling.

D Bangla Errors (QA Prompt)

Table 8 presents representative Bangla error cases for the QA-style prompt. A recurring pattern is that the model often identifies the correct medical concept but fails to reproduce the exact mention boundaries, especially for multiword symptoms and colloquial expressions, resulting in **Boundary Mismatch**. We also observe **Hallucinations** when generic medical advice or common words are treated as symptoms/conditions even when the gold annotation is empty. In contrast, **All-Missed** cases reflect severe recall failures where the model returns an empty JSON despite clear entity cues (e.g., explicit age or specialist mentions). Finally, **Type Confusion** commonly arises between semantically adjacent categories (e.g., tests/procedures vs. conditions, or informal medication names vs. remedies), indicating that the model recognizes the span but struggles with fine-grained label assignment.

E Basque Errors (QA Prompt)

Table 9 shows representative Basque error cases for the same QA-style prompt. Across prompts, the dominant failure mode is **Hallucinations**, suggesting that the model is prone to over-generation in Basque medical text: short phrases and anatomical references are frequently over-labeled as medical entities even when the gold annotation contains none. We also see **Boundary Mismatch** in cases where the model selects a shorter head noun rather than the full descriptive mention, which is penalized under exact-match evaluation. Compared with Bangla, Basque exhibits more frequent over-extraction and fewer purely recall-only failures, consistent with the overall error distribution where FP-heavy errors occupy a larger fraction of error-only cases.

Hierarchy-Aware Hyperbolic and Semantic Reranking for Ontology-Based Phenotype Linking

Thomas Labbé[1,2], Moussa Baddour[2], Axel Bonesteve[3,4], Paul Rollier[3], Marie de Tayrac[2,3,4], Olivier Dameron[2,5]

[1]Orange Research [2]b<>com
[3]Univ Rennes, Department of Molecular Genetics and Genomics, CHU Rennes, IGDR-UMR6290, CNRS
[4]Geeng genetic Genius [5]Univ Rennes, Inria, CNRS, IRISA - UMR 6074
thomas.labbe@gmail.com

Abstract

Extracting structured knowledge from unstructured text is a fundamental challenge in machine learning, particularly for concepts organized within complex hierarchical ontologies. In genomics, identifying phenotypes from clinical narratives is crucial for diagnostic precision, yet current methods struggle with contextual interpretation and subtle clinical descriptions. We present a hierarchy-aware workflow for ontology-based phenotype linking that combines semantic and hierarchical signals. Our approach integrates Large Language Models for span detection with retrieval and a hybrid reranking strategy using both Euclidean (semantic) and hyperbolic (hierarchical) embeddings trained on the Human Phenotype Ontology. We show that while hyperbolic embeddings alone do not outperform standard semantic retrieval, they provide complementary structural signals that improve performance over strong baselines when combined with Euclidean representations. In particular, the hybrid approach outperforms existing state-of-the-art methods and yields more hierarchically coherent predictions, especially in settings involving implicit phenotype mentions. Experiments on a public benchmark (ID-68) and a newly released clinical dataset (CHU-50), publicly released with code and data [1], highlight both performance gains and improved alignment with ontology structure. We further introduce a hierarchy-aware evaluation framework that reflects clinical relevance beyond exact-match metrics.

1 Introduction

Extracting structured knowledge from unstructured text is a central challenge in machine learning, particularly when the target concepts are organized within complex hierarchical ontologies (Figure 1). Phenotypes, as observable traits linking clinical observations to genetic conditions, play a crucial role in diagnosis, treatment planning, and biomedical research Son et al. (2018); Yuan et al. (2022); Mao et al. (2025). While recent advances in natural language processing have improved the extraction of explicitly stated phenotypes, existing systems often struggle to identify implicit or context-dependent mentions.

Current approaches Feng et al. (2023); Luo et al. (2021); Arbabi et al. (2019) frequently rely on flat embedding spaces, which are inadequate for modeling the hierarchical relationships intrinsic to phenotypic ontologies such as the Human Phenotype Ontology (HPO) Robinson et al. (2008). In contrast, hyperbolic geometry naturally models hierarchical (tree-like) structures Gromov (1987) and has been underexplored within this domain. Furthermore, retrieval-based systems are often constrained by their reliance on exact matches or shallow semantic representations. We also argue that existing evaluation metrics widely used in the field (Groza et al., 2024) present further limitations: in practice, clinicians may interpret phenotype mentions differently, as no individual possesses exhaustive knowledge of HPO or applies it uniformly. Consequently, a single reference can yield multiple, equally valid annotations. This variability challenges single-label prediction paradigms and motivates hierarchy-aware approaches, where related terms (e.g., ancestors of a target phenotype), are considered clinically meaningful, albeit less specific.

In this paper, our contributions are four-fold:

i. We propose a workflow for phenotype extraction from clinical text that integrates Large Language Models (LLMs) with retrieval and hierarchical reranking using hyperbolic embeddings trained on HPO.

ii. We introduce a hierarchy-aware evaluation framework that leverages the target ontology structure to provide a more nuanced and clin-

[1] https://github.com/labbeth/HyperRAG/

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 45–63
July 3-4, 2026 ©2026 Association for Computational Linguistics

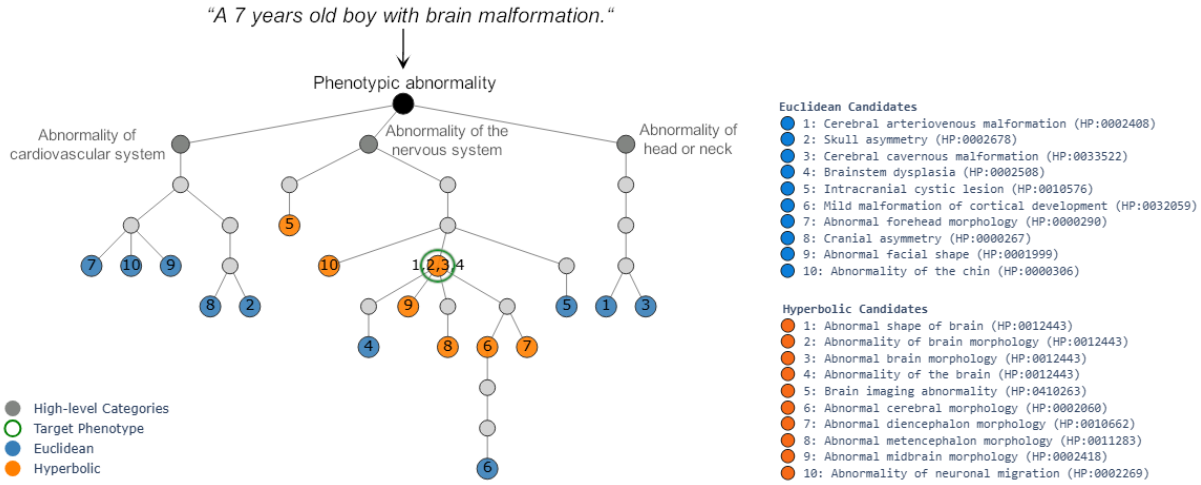


Figure 1: Ontology-based entity linking illustration. Given a clinical sentence, *Top-10* candidates from Euclidean (blue) and Hyperbolic (orange) reranking are overlaid on HPO. Hyperbolic-reranked candidates cluster within the correct branch with lower average distance to target (green), demonstrating improved hierarchical alignment.

ically relevant assessment of extraction systems.

iii. We demonstrate the effectiveness of our approach through comprehensive experiments on both benchmark and challenging real-world datasets, showing consistent improvements, particularly in scenarios with implicit entity mentions.

iv. We make all generated training and evaluation datasets publicly available through our repository, and will release trained models to support reproducibility and further research in ontology-based entity linking.

As part of our workflow validation, we also empirically confirm that hyperbolic embeddings effectively capture the hierarchical structure of HPO, supporting their use in downstream phenotype linking tasks. Overall, our ontology-aware workflow offers a robust solution for phenotype extraction from text, with the potential to enhance diagnostic accuracy and advance biomedical NLP.

2 Background

The introduction of ontologies such as the HPO has provided a structured framework for organizing phenotypic information and has become the primary target for entity linking in this domain. Early work utilized rule-based heuristics Aronson and Lang (2010); Jonquet et al. (2009); Deisseroth et al. (2019), while more recent studies have adopted transformer-based architectures to extract phenotype mentions directly from text Luo et al. (2021); Feng et al. (2023); Yang et al. (2024). Although improvements have been effective with such approaches, they remain complex and often struggle when phenotype references are implicit Baddour et al. (2024). In particular, full LLM approaches are prone to hallucination issues in mapping phenotype labels to HPO ids Labbé et al. (2023). Emerging retrieval-based paradigms Lewis et al. (2020) offer a promising avenue for addressing some of these challenges by efficiently narrowing the candidate space and mitigating hallucinations. However, this approach has not yet been widely adopted in phenotype extraction pipelines, and its performance in this context remains underexplored.

While ontologies facilitate annotation and retrieval, their hierarchical complexity poses significant challenges for NLP systems. Nickel and Kiela (2017) highlighted the limitations of flat embedding spaces in adequately representing such hierarchical structures. Related works Sala et al. (2018); Sinha et al. (2024); Tifrea et al. (2018) proposed to train hyperbolic embeddings that provide a compelling alternative, as hyperbolic spaces are well-suited for modeling hierarchical relationships, allowing embeddings to more accurately reflect the subsumption structure inherent in ontologies.

The motivation behind our proposed workflow stems from recognizing significant limitations in current phenotype extraction systems. While classical dense retrieval approaches Guo et al. (2024) are effective at retrieving candidates based on general semantic similarity, they fall short in capturing the hierarchical relationships and intricate depen-

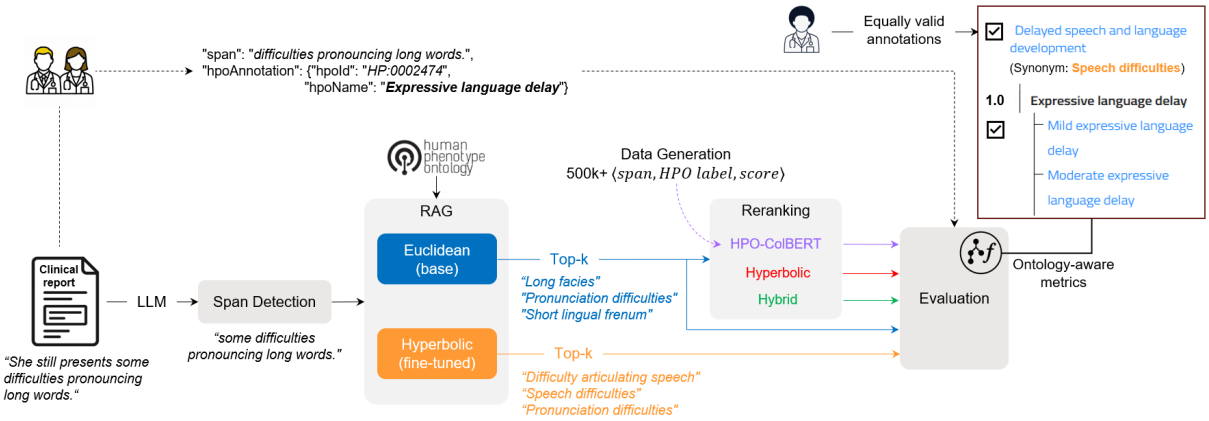


Figure 2: General Workflow. From a clinical report, a LLM first detects spans related to phenotypes. Dense retrieval with Euclidean and Hyperbolic embeddings outputs *Top-k* candidate terms, followed by several reranking strategies. Evaluation uses both clinician-provided ground truth annotations and the HPO ontology structure to account for hierarchical relationships and semantic similarity, recognizing that multiple valid references may exist.

dencies inherent in ontologies such as HPO Huang et al. (2025); Gao et al. (2023). This limitation becomes even more pronounced when dealing with implicit phenotypes not explicitly stated in clinical text, where leveraging ontological relationships can be crucial for accurate identification and resolution Peng et al. (2024).

By integrating hyperbolic embeddings (which naturally encode hierarchical structures) with a reranking mechanism, our workflow bridges the gap between general semantic relevance and ontological hierarchy. This dual approach ensures not only accurate retrieval of phenotypes but also a ranking that reflects their hierarchical significance, providing a comprehensive solution to the limitations of current methods.

3 Methodology

The proposed workflow is illustrated in Figure 2: Given clinical reports, the process consists of four main steps: span detection using an LLM, dense candidate retrieval, reranking of candidates, and evaluation with both standard and ontology-aware metrics.

3.1 Span Identification

We use a pretrained LLM to identify phenotype spans, effectively capturing implicit mentions that may be overlooked by traditional methods. Notably, Baddour et al. (2024) demonstrated that employing an LLM as a span detector outperforms the biomedical Stanza pipeline (Zhang et al., 2021).

3.2 Retrieval

We compute dense span embeddings using either a base or HPO-fine-tuned hyperbolic model. *Top-k* phenotype candidates are retrieved from the HPO ontology based on *cosine similarity* (euclidean model) or normalized *hyperbolic distance* (hyperbolic model).

3.3 Reranking

For each span, the *Top-k* HPO candidates retrieved through dense Euclidean embeddings similarity are reranked using two families of methods: late-interaction and hyperbolic.

Late-interaction reranking

As a strong classical baseline, we fine-tuned a late-interaction model Santhanam et al. (2021) for reranking. While cross-encoders are effective, they are computationally demanding and less adaptable to incorporating soft signals like distance-based scores Jha et al. (2024). Late-interaction models strike a balance by retaining token-level embeddings and applying a late matching function, thus preserving fine-grained information often lost in bi-encoders. This makes them particularly suitable for our reranking task, which involves short text spans and specific target labels.

Hyperbolic reranking

- **Full hyperbolic reranking:** Both the input span and the *Top-k* candidates from the Euclidean retriever are embedded in hyperbolic space. Candidates are then reordered based on their normalized hyperbolic distances to the

input span. We include this setting primarily as an ablation to isolate the effect of purely hierarchy-driven scoring.

- **Hybrid reranking:** This approach combines the cosine similarity between Euclidean embeddings and the hyperbolic distance between hyperbolic embeddings using a weighted sum. Cosine similarity emphasizes semantic closeness, while hyperbolic distance prioritizes candidates with closer hierarchical relationships to the input span. given an input $span$ and a set of candidates $\{C_i\}_{i=1}^{k}$, the hybrid score of a candidate C_i is computed as follows.

$$S_{\text{hybrid}}(C_i, span) = \gamma \cdot S_{\cos} - (1-\gamma) \cdot \hat{d}_{\mathbb{H}} \quad (1)$$

where $S_{\cos}$ is the cosine similarity between C_i and $span$, $\hat{d}_{\mathbb{H}}$ is their normalized hyperbolic distance, and γ is the weighting parameter between the two metrics (see Appendix E for ablation).

Full implementation details (including model architectures, training hyperparameters, normalization strategies, synonym mapping, and retrieval settings) are provided in Appendix A, B and D to facilitate reproducibility. Code and data are publicly available through our repository.

4 Dataset

4.1 Ontologies

The Human Phenotype Ontology (HPO) Robinson et al. (2008) serves as the foundation for our hierarchical embeddings, comprising over 19,000 hierarchically organized phenotypic terms and an extensive set of synonyms.

Additionally, we leverage the SNOMED ontology El-Sappagh et al. (2018) indirectly through a pretrained hyperbolic model (fine-tuned from the same base model). It allows us to assess the relative benefits of utilizing this broad, general-purpose medical ontology in comparison to a highly specialized ontology (HPO) for phenotype extraction tasks.

4.2 Training Data

Hyperbolic Training: Hyperbolic embeddings were trained on HPO triplets extracted from the ontology, with synonym augmentation and filtering to ensure balanced training. We used Hierarchy Transformers He et al. (2024) to effectively capture hierarchical relationships.

Late-interaction Training: We fine-tuned ColBERTv2 Santhanam et al. (2021) on triplets $\langle$*span, HPO label, score*$\rangle$, using a dataset generated by prompting ChatGPT-4o-mini to create diverse clinical sentences for each HPO term. Sentence quality was assessed by clinicians, and further validated using an LLM-as-a-judge approach. After heuristic filtering, we obtained 91,760 high-quality spans, which were used to construct positive and negative pairs, resulting in 510,371 training triplets.

Full data construction details and training parameters for both hyperbolic and late-interaction models are provided in Appendix B and D. Prompts for LLM-as-a-judge are provided in Appendix I.

4.3 Evaluation Dataset

We evaluate our workflow using two datasets. **ID-68** Anazi et al. (2017), a widely used benchmark for phenotype extraction. **CHU-50**, a dataset of 50 fully anonymized clinical notes from a partner hospital, which we publicly release trough our repository. It contains 971 phenotype annotations, approximately 30% of which are implicit. Results are compared to PhenoBERT Feng et al. (2023), the most advanced open-source state-of-the-art solution available.

5 Evaluation

We first evaluated the hyperbolic model independently, prior to conducting the main phenotype extraction experiments.

5.1 Hyperbolic Inner Evaluation

To assess the consistency of the hyperbolic model, we compared its normalized distance metrics with those of the baseline Euclidean model. Specifically, we examined one-hop and multi-hop distances to evaluate the model's ability to capture hierarchical relationships, as well as distances between synonyms and negative pairs to determine whether semantic consistency is preserved.

Additionally, we introduce a *hierarchical representation power* plot to visualize the model's capacity to encode hierarchy while maintaining semantic coherence. This radar chart displays the average distances for one-hop, multi-hop, and synonym pairs, alongside the inverse average distance for negative pairs. This visualization enables us

to assess whether the embedding space has been structured as intended.

5.2 Phenotypes Linking Evaluation

In practice, generating a comprehensive list of phenotypes for each patient is crucial for accurate diagnosis, making recall-based metrics (*recall@k* and *miss_rate@k*) the primary focus. While *Top-1* precision is reported for comparison with existing single-output methods (e.g. PhenoBERT), it fails to reflect the clinical relevance of alternative valid annotations. To further assess ranking quality, we include Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG).

However, these traditional metrics are limited when based solely on exact matches, which is the prevailing evaluation paradigm in current solutions. In practice, a parent term of the target phenotype often conveys relevant information, even if it is less specific, and predictions involving descendants or related terms should not be considered entirely incorrect.

To address this limitation, we introduce a novel hierarchical evaluation framework that leverages the structure of HPO to weight candidate scores according to their proximity to the ground truth. These *relationship scores* are computed based on the specific type of relationship between the candidate $\mathcal{C}$ and the target phenotype $\mathcal{T}$.

Direct relationships

$$w_{direct}(\mathcal{C},\mathcal{T}) = \begin{cases} \frac{\alpha}{p\times(1+|d|)}, & p>0 \\ 1, & p=0 \end{cases} \quad (4)$$

where α is a constant factor, p is the number of ancestors/descendants between $\mathcal{C}$ and $\mathcal{T}$, and d is the distance between $\mathcal{C}$ and $\mathcal{T}$.

Indirect relationship

$$w_{indirect} = \frac{\beta}{c\times(1+d_l)} \quad (5)$$

where β is a constant factor, c is the number of immediate descendants of the most specific common ancestor between $\mathcal{C}$ and $\mathcal{T}$, and d_l is the distance between $\mathcal{C}$ and the farthest HPO leaf.

See Appendix E for α and β settings.

By combining absolute distances with the cardinality of surrounding phenotypes, these functions effectively characterize the strength of relationships between HPO terms, balancing both proximity and semantic relevance. Throughout this paper, the term *weighted* refers to evaluation metrics that incorporate these hierarchical weightings.

In addition, we introduce specific metrics to assess how well the models respect the ontology's structure: the average number of hops between each candidate and the target phenotype; the average branch coverage, defined as the proportion of candidates within the same branch as the target; and the distribution of relationship types by position, measuring the proportions of exact matches, ancestors, descendants, cousins, or candidates with no direct path to the target. We also report the proportion of close candidates, defined as those with a *relationship score* above a specified threshold.

6 Results

6.1 Hyperbolic consistency

Figure 3. presents the distributions of one-hop and multi-hop distances for both the Euclidean model and the fine-tuned hyperbolic model. The distributions for the hyperbolic model are notably narrower and exhibit lower means, particularly for multi-hop distances, indicating a more faithful representation of the ontology's hierarchical structure.

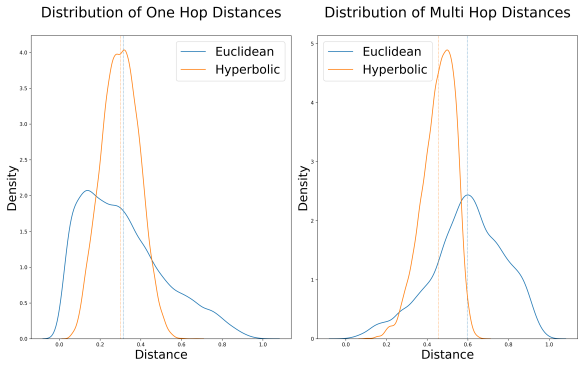


Figure 3: Euclidean (blue) vs Hyperbolic (orange) distance distributions between one-hop (left) and multi-hop (right) phenotypes in the Human Phenotypes Ontology (HPO), computed on the test set (10%). Vertical lines represent respective means.

Furthermore, the resulting hyperbolic model preserves the semantic structure of the base model, as illustrated in Figure 4. Although the average distance between negative pairs is slightly reduced, these pairs remain well separated from positive examples. Notably, synonyms within the HPO are now positioned closer together, and multi-hop

phenotypes are significantly closer than in the Euclidean embedding space, reflecting improved hierarchical modeling. In contrast, one-hop phenotypes are only marginally closer, which is expected given the typically strong semantic similarity between such terms (e.g.: *Iris coloboma* is semantically closer to its one-hop parent *Coloboma* than the 2-hops *Abnormal eye morphology*).

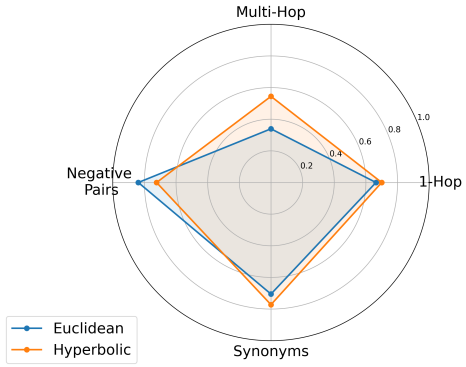


Figure 4: Semantic and Hierarchical Representation Power. Values represent normalized similarity (or inverse for negative pairs) between one-hop, multi-hop, and synonyms in Euclidean and Hyperbolic spaces.

6.2 Phenotypes Linking

For the hybrid approach, we conducted evaluations with several values of γ (reported in Appendix E). Since the performance differences were marginal, we selected $\gamma = 0.5$ for its simplicity and interpretability, recognizing that this value may not be optimal but provides a balanced influence of both models. For results reported in this section, statistical significance was assessed using paired bootstrap resampling at the report level (Appendix C).

Retrieval

As shown in Figure 5, the hyperbolic retriever model underperforms compared to other methods on both datasets, with recall further decreasing when hyperbolic reranking is applied to Euclidean retriever candidates.

This behavior reflects a trade-off between hierarchical organization and local semantic similarity: the Poincaré geometry emphasizes vertical ontological relations, while training primarily on *is-a* links and the mismatch between clinical spans and ontology-level supervision may further limit fine-grained semantic alignment, motivating the use of hyperbolic representations as a complementary rather than standalone signal.

For ID-68 (top plots), hybrid reranking improves recall from $k = 5$ onwards, and late-interaction reranking becomes effective from $k = 15$, though it does not surpass the hybrid approach. Both Euclidean and hybrid reranking outperform prior systems in recall from $k = 3$ onwards and set new state-of-the-art at $k = 1$ in the weighted setting ($+9$), with gains up to $+18$ at $k = 15$. Hybrid reranking also achieves the lowest miss rate, reducing misses by 17 at $k = 15$.

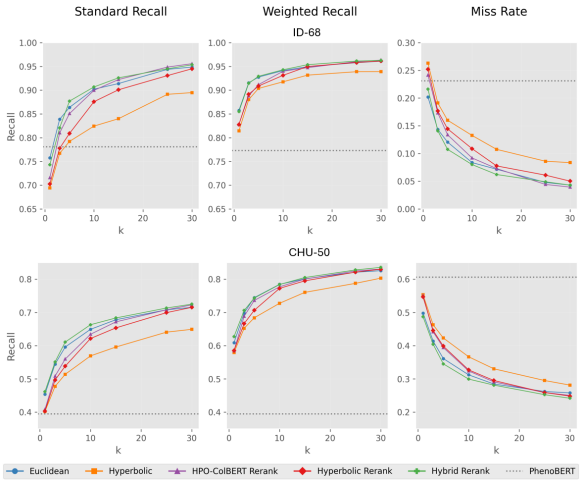


Figure 5: Recall and miss rate on the ID-68 (top) and CHU-50 (bottom) datasets. Standard (exact match) and weighted recall (considering HPO hierarchy) are shown for raw dense retrieval (Euclidean, Hyperbolic) and reranked results (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate indicates the proportion of ground truth phenotypes not retrieved.

On CHU-50 (bottom plots), similar trends are observed, with all models outperforming state-of-the-art PhenoBERT across all three metrics from $k = 1$, achieving a $+23$ increase in recall and an 18-point reduction in miss rate. This is consistent with previous findings Baddour et al. (2024), as PhenoBERT is less effective at capturing implicit phenotype references. Overall, the logarithmic shape of the recall and miss rate curves across both datasets indicates that all models rank correct candidates highly.

Complementary cross-ontology analyses (Appendix H) using a SNOMED-trained hyperbolic model show that, although it underperforms an HPO-trained hyperbolic model on HPO-defined targets (as expected), the SNOMED component can inject broader hierarchical cues that yield

a slight recall gain on ID-68 and a lower miss rate on CHU-50. These effects suggest that non–domain-specialized embeddings can add complementary structure in difficult cases, but the gains are modest and do not change the observation that HPO-centric retrieval and hybrid reranking remain the most effective overall, with the Euclidean baseline strong and hierarchy-aware metrics further favoring hyperbolic components.

Ranking

Tables 1 and 2 present weighted MRR and NDCG results for ID-68 and CHU-50. On ID-68, the Euclidean and Hybrid Rerank models achieve the highest MRR, indicating top-ranked correct phenotypes. The Hyperbolic model performs slightly lower, but the gap narrows with hierarchy-aware metrics, highlighting its strength in capturing ontological relationships. NDCG scores are also high across all models, with Euclidean and Hybrid Rerank exceeding 0.94 at $k = 1$. As k increases, both metrics decrease slightly, but Hybrid Rerank consistently maintains strong performance.

Table 1: Weighted Metrics by Model and k for ID-68

Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Weighted MRR					
Euclidean	**0.857**	**0.882**	**0.884**	**0.885**	**0.885**
Hyperbolic	0.814	0.841	0.845	0.847	0.848
HPO-ColBERT Rerank	0.828	0.851	0.855	0.858	0.858
Hyperbolic Rerank	0.828	0.853	0.857	0.859	0.860
Hybrid Rerank	**0.855**	**0.881**	**0.883**	**0.885**	**0.885**
Weighted NDCG					
Euclidean	**0.949**	**0.936**	**0.923**	**0.901**	0.883
Hyperbolic	0.932	0.922	0.902	0.877	0.861
HPO-ColBERT Rerank	0.931	0.928	0.912	0.886	0.870
Hyperbolic Rerank	0.944	0.927	0.912	0.884	0.870
Hybrid Rerank	0.943	**0.936**	**0.924**	**0.904**	**0.891**

On the more challenging CHU-50 dataset, which contains a higher proportion of implicit phenotype mentions, all models exhibited lower MRR and NDCG scores compared to ID-68. However, the Hybrid Rerank model outperforms others for all k values.

These results indicate that the hybrid approach, which combines semantic similarity from Euclidean embeddings with hierarchical proximity from hyperbolic embeddings, is particularly effective in ranking the most relevant phenotypes at the top, even in complex, real-world clinical text. The consistently strong MRR and NDCG scores for the Hybrid Rerank model confirm that combining semantic and hierarchical signals yields superior candidate ranking. Hierarchy-aware

Table 2: Weighted Metrics by Model and k for CHU-50

Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Weighted MRR					
Euclidean	0.631	0.659	0.667	0.670	0.671
Hyperbolic	0.588	0.613	0.619	0.625	0.627
HPO-ColBERT Rerank	0.585	0.624	0.633	0.636	0.638
Hyperbolic Rerank	0.597	0.629	0.637	0.643	0.644
Hybrid Rerank	**0.647**	**0.669**	**0.676**	**0.680**	**0.681**
Weighted NDCG					
Euclidean	0.841	**0.853**	**0.832**	**0.807**	0.792
Hyperbolic	0.834	0.818	0.796	0.772	0.751
HPO-ColBERT Rerank	0.844	0.821	0.805	0.783	0.768
Hyperbolic Rerank	0.851	0.841	0.815	0.787	0.779
Hybrid Rerank	**0.868**	0.848	**0.835**	**0.806**	**0.800**

weighted metrics further demonstrate the value of hyperbolic embeddings in capturing nuanced ontological relationships, especially when exact matches are unavailable but related terms remain clinically relevant.

Overall, hybrid reranking yields statistically significant improvements in weighted recall@1 and weighted MRR on the CHU-50 dataset, while maintaining performance comparable to Euclidean retrieval on ID-68 (detailed statistical significance results are provided in Appendix C).

6.3 Ontology-based Metrics

While hyperbolic models may not always outperform Euclidean models at top ranks, they show greater robustness as the candidate list grows, maintaining lower average hops and higher branch coverage (Figure 6).

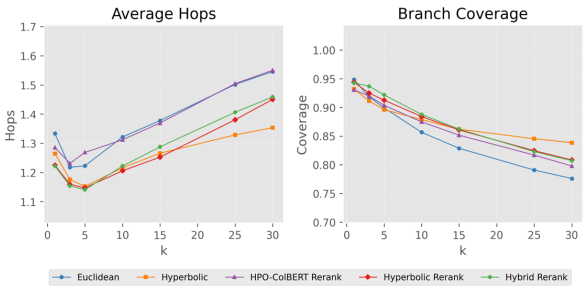


Figure 6: Ontology-based metrics on the ID-68 dataset across *Top-k* candidates. Average number of hops between the candidates and the target phenotypes (left), and proportion of candidates within the ontology branch of the target phenotypes (right).

This observation is further supported by a detailed analysis of ontological relationships (Appendix F). Figure 7 shows a high percentage of exact matches at *Top-1*, confirming the pipeline's effectiveness. However, deeper analysis reveals important distinctions between modeling approaches.

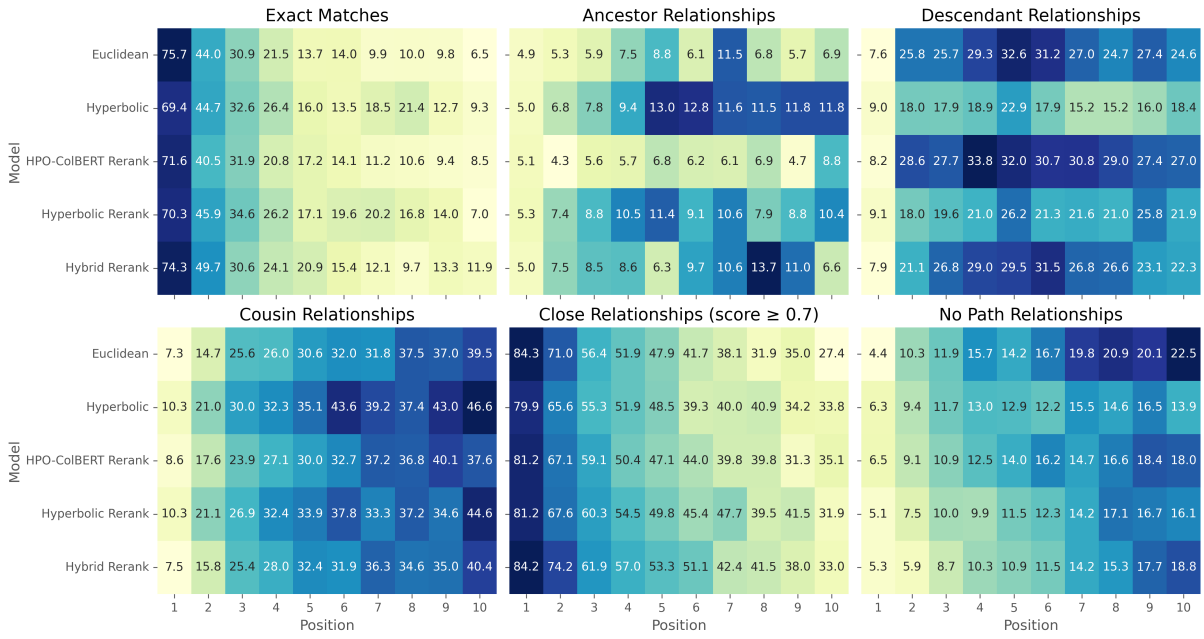


Figure 7: Distribution of relationship types by position and model for *Top-10* candidates ($k = 10$) on ID-68 dataset. The heatmaps show percentages for six relationship categories (Appendix F) between predicted and ground truth phenotypes. For closeness relationships, only those with scores above 0.7 (Equations (4) and (5)) are considered. Rows correspond to different models, and columns represent candidate positions ranked from 1 to 10. Higher values indicate stronger presence of the relationship type at the given rank and model.

Hyperbolic models (both raw output and reranking) exhibit significantly higher proportions of ancestor and cousin relationships, while showing fewer descendant relationships compared to Euclidean or HPO-ColBERT models. This pattern strongly suggests that hyperbolic approaches better capture the hierarchical structure of the HPO ontology in both vertical and horizontal dimensions. The tendency to "move upward" in the hierarchy toward more general terms rather than "downward" toward more specific ones aligns with theoretical expectations of hyperbolic geometry, where distances increase exponentially with depth in the hierarchy.

Notably, hyperbolic models maintain semantic relevance at higher ranks, preserving close relationships and yielding fewer unrelated candidates as k increases. This semantic consistency at higher ranks has important implications for clinical applications, as it reduces the risk of missing relevant phenotypes (false negatives) when examining a broader set of candidates. The hybrid reranking approach combines the strengths of both geometries, achieving strong exact matching at top positions and semantic coherence at higher ranks. This balanced performance confirms the value of integrating both approaches for optimal phenotype retrieval in clinical settings. Similar trends are observed on the CHU-50 dataset (Figure 16 in Appendix H).

7 Conclusion

In summary, we introduce a hierarchy-aware workflow combining LLM-based span detection, dense retrieval, and hybrid reranking using hyperbolic embeddings for phenotype linking. Experiments on benchmark and real-world datasets show improvements in recall, miss rate, and ranking quality, alongside better hierarchical consistency, particularly under hierarchy-aware metrics. The hybrid reranking strategy, integrating semantic and ontological signals, consistently outperforms prior systems, especially for implicit phenotype mentions. These results highlight the complementary roles of Euclidean and hyperbolic embeddings, capturing local semantic similarity and global ontology structure, and yielding more clinically meaningful and interpretable predictions. Finally, the proposed evaluation framework and released datasets support more nuanced and clinically relevant assessment of phenotype extraction systems. This approach can be extended to other domains involving hierarchical ontologies.

Limitations

While our approach demonstrates notable improvements, several limitations remain. Although hyperbolic embeddings effectively capture hierarchical relationships within the ontology, they may fail to fully represent implicit semantic nuances in clinical text. Future work could explore multi-task training strategies jointly modeling hierarchy and semantic similarity.

Another limitation concerns data sparsity: reliance on annotated datasets may restrict generalizability to unseen or rare phenotypes. Additionally, hyperbolic distances are not always meaningful for all term pairs, suggesting that further refinement of the hyperbolic modeling framework is needed.

Although CHU-50 contains a substantial proportion of implicit phenotype mentions, the current version of the dataset does not explicitly distinguish implicit and explicit annotations. Consequently, evaluations are aggregated across all mention types. Future versions of the dataset will include dedicated annotation flags to enable finer-grained analyses of implicit phenotype normalization performance.

Regarding recall performance, our analysis did not reveal clear patterns explaining why SNOMED embeddings occasionally outperform HPO embeddings. While broader-domain supervision may provide complementary signals in some cases, SNOMED embeddings underperform HPO-based models on the CHU-50 dataset, indicating that domain-specific hierarchical information remains crucial for complex clinical text.

More generally, the effectiveness of hyperbolic embeddings depends on the underlying ontology structure. Ontologies with complex or transversal relationships may limit the benefits of hyperbolic geometry, potentially requiring alternative training strategies or geometric formulations. In this respect, extending our workflow to other hierarchical coding systems such as the International Classification of Diseases (ICD) Blanco et al. (2020); Zhou et al. (2021) would require additional investigation, as its granularity, structural conventions, and limited mappings to HPO Tan et al. (2024) may constrain direct transfer.

Our approach also introduces additional system components compared to simpler pipelines such as standalone LLM inference or NER+EL models. However, most of the added cost is incurred offline, as hyperbolic embeddings are trained once and reranking relies on precomputed representations without additional LLM calls at inference time.

Finally, due to computational constraints, we did not perform an exhaustive search over hybrid reranking parameters (although we explored different values of γ in Appendix E) or evaluation weighting schemes, nor did we investigate alternative embedding backbones, particularly those fine-tuned on medical or clinical data, which could further improve performance.

Ethical Considerations

The CHU-50 dataset introduced in this work consists of clinical notes with phenotype annotations and is used exclusively for evaluation purposes. No real clinical data are used for model training. All notes were manually generated by clinicians based on real clinical reports, but the content is fully original and anonymized. No personal data are present in the dataset, and care was taken to ensure that no combination of symptoms could indirectly identify any individual. This design ensures compliance with ethical standards and protects patient privacy.

Additionally, the proposed workflow relies in part on probabilistic components, including large language models used for span detection. These outputs should be interpreted with caution. If used in a medical context, all results must be reviewed and validated by qualified healthcare professionals, and the system should not be used for autonomous clinical decision-making.

References

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T Asi, Saud Alsahli, Amal Alhashem, Hanan E Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, and 1 others. 2017. Expanding the genetic heterogeneity of intellectual disability. Human genetics, 136:1419–1429.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, and 1 others. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. JMIR medical informatics, 7(2):e12596.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236.

Moussa Baddour, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, and Thomas Labbé. 2024. Phenotypes extraction from text: Analysis and perspective in the llm era. In 2024

IEEE 12th International Conference on Intelligent Systems (IS), pages 1–8. IEEE.

Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label icd classification: Sensitivity to hospital service and time. IEEE Access, 8:183534–183545.

Cole A Deisseroth, Johannes Birgmeier, Ethan E Bodle, Jennefer N Kohler, Dena R Matalon, Yelena Nazarenko, Casie A Genetti, Catherine A Brownstein, Klaus Schmitz-Abe, Kelly Schoch, and 1 others. 2019. Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine, 21(7):1585–1593.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. arXiv preprint arXiv:2401.08281.

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. BMC medical informatics and decision making, 18:1–19.

Yuhao Feng, Lei Qi, and Weidong Tian. 2023. Phenobert: A combined deep learning method for automated recognition of human phenotype ontology. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(2):1269–1277.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).

Mikhael Gromov. 1987. Hyperbolic groups. In Essays in group theory, pages 75–263. Springer.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. 2024. An evaluation of gpt models for phenotype concept recognition. BMC Medical Informatics and Decision Making, 24(1):30.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. 2024. Language models as hierarchy encoders. In Advances in Neural Information Processing Systems, volume 37, pages 14690–14711. Curran Associates, Inc.

Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-augmented generation with hierarchical knowledge. arXiv preprint arXiv:2503.10150.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. 2024. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. arXiv preprint arXiv:2408.16672.

Clement Jonquet, Nigam H Shah, Cherie H Youn, Mark A Musen, Chris Callendar, and Margaret-Anne Storey. 2009. Ncbo annotator: semantic annotation of biomedical data. In ISWC 2009-8th International Semantic Web Conference, Poster and Demo Session, 171.

Thomas Labbé, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. 2023. Chatgpt for phenotypes extraction: one model to rule them all? In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 1–4. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. 2021. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. Bioinformatics, 37(13):1884–1890.

Xiaohao Mao, Yu Huang, Ye Jin, Lun Wang, Xuanzhong Chen, Honghong Liu, Xinglin Yang, Haopeng Xu, Xiaodong Luan, Ying Xiao, and 1 others. 2025. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. npj Digital Medicine, 8(1):68.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems, 30.

OpenAI. 2023. Chatgpt (mar 23 version). Https://chat.openai.com/chat.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics, 83(5):610–615.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In International conference on machine learning, pages 4460–4469. PMLR.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. 2024. Learning structured representations with hyperbolic embeddings. Advances in Neural Information Processing Systems, 37:91220–91259.

Jung Hoon Son, Gangcai Xie, Chi Yuan, Lyudmila Ena, Ziran Li, Andrew Goldstein, Lulin Huang, Liwei Wang, Feichen Shen, Hongfang Liu, and 1 others. 2018. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. The American Journal of Human Genetics, 103(1):58–73.

Amelia LM Tan, Rafael S Gonçalves, William Yuan, Gabriel A Brat, Robert Gentleman, Isaac S Kohane, Aaron J Masino, Adeline Makoudjou, Adem Albayrak, Alba Gutiérrez-Sacristán, and 1 others. 2024. Implications of mappings between international classification of diseases clinical diagnosis codes and human phenotype ontology terms. JAMIA open, 7(4).

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincar\'e glove: Hyperbolic word embeddings. arXiv preprint arXiv:1810.06546.

Wei F. Dong L. Bao H. Yang N. Zhou M. Wang, Y. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In NeurIPS.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. 2024. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. Patterns, 5(1).

Xiao Yuan, Jing Wang, Bing Dai, Yanfang Sun, Keke Zhang, Fangfang Chen, Qian Peng, Yixuan Huang, Xinlei Zhang, Junru Chen, and 1 others. 2022. Evaluation of phenotype-driven gene prioritization methods for mendelian diseases. Briefings in Bioinformatics, 23(2):bbac019.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. Journal of the American Medical Informatics Association, 28(9):1892–1899.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5948–5957.

A Implementation Details

Span detection: for consistency with Baddour et al. (2024) and comprehensive coverage, we utilized ChatGPT-3.5 model OpenAI (2023). Corresponding prompt is described in Appendix I.

Embeddings: we used *all-MiniLM-L12-v2* Wang (2020) as the embeddings base model, from which the hyperbolic model was fine-tuned. We set *k=30* to substantially reduce the candidate space while still allowing for meaningful reranking improvements.

Candidate retrieval: FAISS Douze et al. (2024) is employed as the vector store and *Top-k* retriever for the Euclidean model. In contrast, the hyperbolic model utilizes a dedicated vector index and retrieval mechanism.

To ensure consistency in distance measurements across experiments, we normalize the hyperbolic distances in the Poincaré ball using a global normalization strategy (2):

$$\hat{d}_{\mathbb{H}}(u,v) = \frac{d_{\mathbb{H}}(u,v)}{\max_{p,q\in \text{HPO}} d_{\mathbb{H}}(p,q)} \quad (2)$$

where $d_{\mathbb{H}}(u,v)$ is the hyperbolic distance between terms u and v in the hyperbolic space $\mathbb{H}$, $\hat{d}_{\mathbb{H}}$ is the normalized hyperbolic distance, and $\max_{p,q\in \text{HPO}} d_{\mathbb{H}}(p,q)$ is the maximum hyperbolic distance between two terms in the HPO ontology.

Synonyms in the dense retrieval output are mapped to their original HPO terms using a precomputed synonym-to-ID mapping. This ensures consistency in distance calculations throughout the workflow.

Hybrid reranking: given an input $span$ and a set of candidates $\{C_i\}_{i=1}^{k}$, the hybrid score of a candidate C_i is computed as follows.

$$S_{\text{hybrid}}(C_i, span) = \gamma \cdot S_{\cos} - (1-\gamma)\cdot \hat{d}_{\mathbb{H}} \quad (3)$$

where $S_{\cos}$ is the cosine similarity between C_i and $span$, $\hat{d}_{\mathbb{H}}$ is their normalized hyperbolic distance, and γ is the weighting parameter between the two metrics.

B Dataset Construction

Hyperbolic Training Data

Hierarchical relationships are extracted from the HPO OWL file using DeepOnto and the ELK

reasoner. Following the methodology of He et al. (2024), we generated a dataset of triplets ⟨*child, parent, label*⟩, where the label is a binary indicator of a positive or negative example, and triplets ⟨*child, parent, negative*⟩, where the negative term is not a parent of the child. We used random negative sampling strategy in this implementation. Given that most HPO phenotypes are associated with multiple synonyms, we augmented the dataset by including all possible synonym combinations within each triplet. This augmentation enhances the robustness of the resulting embeddings to varied term formulations. To prevent excessive class imbalance, we applied a filtering strategy, limiting each synonym to a maximum of five occurrences.

Name	Definition
`WHITELIST_WORDS`	Set of medically relevant or specific terms; if any are present in a span, the span is always kept.
`BLACKLIST_PATTERNS`	Set of phrases indicating non-informative or undesirable content; if present (and not overridden by whitelist), the span is always filtered out.
`GENERIC_WORDS`	Set of common, non-specific words; used to penalize spans that are mostly generic.
`VAGUE_PATTERNS`	Set of vague or non-specific phrases; presence reduces the span's score.

Table 3: Heuristic lists used for span filtering.

Late-interaction Training Data

We fine-tuned the ColBERTv2 model Santhanam et al. (2021) on triplets of the form ⟨*span, HPO label, score*⟩, where the score represents a similarity measure. To construct a comprehensive training dataset, we first used ChatGPT-4o-mini to generate 10 clinical report sentences for each HPO term in the ontology. To ensure diversity and representativeness, we specified requirements for each batch of 10 sentences (e.g., at least two sentences should be implicit, up to two should include measurements, etc.). For this iteration, we excluded cases where a sentence refers to multiple phenotypes. For each generated sentence, we further prompted ChatGPT-4o-mini to extract the most precise span capturing the clinical observation of the target phenotype. This process resulted in the *HPO_sentences_spans* dataset, comprising over 200,000 clinical sentences and corresponding spans, covering the entire set of HPO terms. All prompts are provided in Appendix I for reproducibility.

After qualitative analysis of generated spans, some of them appear to be uninformative (e.g.: for the sentence "*The bladder capacity measured at 150 mL, which is below the normal range.*", the output span is "*below the normal range*" which lacks context to be relevant). We conducted a deep qualitative analysis of the spans with our clinicians in order to define filtering rules. Table 3 define the heuristics elements used in the process. The scoring procedure and the span filtering are detailed in pseudo-code in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 Compute Span Score

Require: span (string)
if span contains any **BLACKLIST_PATTERN then**
  if span contains any **WHITELIST_WORD then**
    return 1.0
  else
    return 0.0
  end if
end if
if span contains any **WHITELIST_WORD then**
  return 1.0
end if
if number of words in span ≥ 3 **then**
  score $\leftarrow$ 1.0
else
  score $\leftarrow$ 0.3
end if
generic_ratio $\leftarrow$ (number of **GENERIC_WORDS** in span) / (total words)
if generic_ratio > 0.6 **then**
  score $\leftarrow$ score -0.5
end if
if span contains any **VAGUE_PATTERN then**
  score $\leftarrow$ score -0.3
end if
return max(score, 0.0)

Algorithm 2 Filter Spans

Require: spans (list of strings)
for each span in spans **do**
  score $\leftarrow$ **Compute Span Score**(span)
  if score ≥ 0.2 **then**
    keep span
  else
    filter out span
  end if
end for

As a result, this lower-quality spans filtering yields 91,760 unique high-quality spans (with 2,167 unique spans filtered out).

Finally, we leveraged the trained hyperbolic model for scoring: positive *(span, HPO label)* pairs from the generated dataset were assigned a score

Table 4: Paired bootstrap confidence intervals (95%) for weighted Recall and weighted MRR on the ID-68 dataset. Δ denotes the mean difference between Hybrid and Euclidean retrieval ($\Delta = B - A$).

k	Metric	Euclidean (A)	Hybrid (B)	Δ [95% CI]
1	Recall	0.8663	0.8584	-0.0079 [-0.0254, $+0.0092$]
3	Recall	0.9239	0.9218	-0.0021 [-0.0113, $+0.0072$]
5	Recall	0.9340	0.9350	$+0.0009$ [-0.0078, $+0.0099$]
10	Recall	0.9452	0.9475	$+0.0024$ [-0.0045, $+0.0089$]
15	Recall	0.9544	0.9572	$+0.0028$ [-0.0016, $+0.0073$]
25	Recall	0.9621	0.9646	$+0.0025$ [-0.0023, $+0.0074$]
30	Recall	0.9643	0.9657	$+0.0014$ [-0.0023, $+0.0049$]
1	MRR	0.8663	0.8584	-0.0079 [-0.0254, $+0.0092$]
3	MRR	0.8901	0.8856	-0.0045 [-0.0165, $+0.0074$]
5	MRR	0.8919	0.8879	-0.0040 [-0.0162, $+0.0079$]
10	MRR	0.8930	0.8892	-0.0038 [-0.0155, $+0.0076$]
15	MRR	0.8931	0.8896	-0.0036 [-0.0151, $+0.0077$]
25	MRR	0.8933	0.8898	-0.0035 [-0.0150, $+0.0078$]
30	MRR	0.8933	0.8898	-0.0035 [-0.0149, $+0.0076$]

Table 5: Paired bootstrap confidence intervals (95%) for weighted Recall and weighted MRR on the CHU-50 dataset. Bold values indicate statistically significant differences (confidence interval excludes zero).

k	Metric	Euclidean (A)	Hybrid (B)	Δ [95% CI]
1	Recall	0.5966	0.6152	$\mathbf{+0.0186}$ [$\mathbf{+0.0033}$, $\mathbf{+0.0345}$]
3	Recall	0.6792	0.6899	$+0.0107$ [-0.0039, $+0.0265$]
5	Recall	0.7215	0.7234	$+0.0019$ [-0.0090, $+0.0131$]
10	Recall	0.7586	0.7592	$+0.0006$ [-0.0100, $+0.0120$]
15	Recall	0.7733	0.7782	$+0.0048$ [-0.0074, $+0.0189$]
25	Recall	0.7897	0.7984	$\mathbf{+0.0088}$ [$\mathbf{+0.0005}$, $\mathbf{+0.0189}$]
30	Recall	0.7946	0.8045	$\mathbf{+0.0099}$ [$\mathbf{+0.0021}$, $\mathbf{+0.0191}$]
1	MRR	0.5966	0.6152	$\mathbf{+0.0186}$ [$\mathbf{+0.0033}$, $\mathbf{+0.0345}$]
3	MRR	0.6279	0.6417	$\mathbf{+0.0138}$ [$\mathbf{+0.0028}$, $\mathbf{+0.0247}$]
5	MRR	0.6359	0.6474	$\mathbf{+0.0115}$ [$\mathbf{+0.0023}$, $\mathbf{+0.0207}$]
10	MRR	0.6391	0.6506	$\mathbf{+0.0115}$ [$\mathbf{+0.0022}$, $\mathbf{+0.0209}$]
15	MRR	0.6399	0.6516	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0210}$]
25	MRR	0.6404	0.6521	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0209}$]
30	MRR	0.6405	0.6522	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0208}$]

of 1, while negative pairs were created by pairing spans with other phenotypes and assigning scores based on the normalized hyperbolic distance to the target phenotype. Both hard negatives (phenotypes within the same branch, up to three hops away) and easy negatives (phenotypes outside the target branch) were included. The final training set consists of 510,371 triplets $\langle$*span, HPO label, score*$\rangle$.

C Statistical Significance Analysis

Statistical significance was assessed using paired bootstrap resampling at the report level (2,000 resamples), with confidence intervals computed on mean differences between hybrid reranking and Euclidean retrieval.

ID-68. On the ID-68 dataset (Table 4), which contains mostly explicit phenotype mentions, hybrid reranking achieves performance comparable to Euclidean retrieval for both weighted recall and weighted MRR across all evaluated values of k. Paired bootstrap confidence intervals consistently include zero, indicating no statistically significant differences between the two approaches. This suggests that integrating hierarchical information does not degrade either retrieval coverage or ranking quality in settings where semantic similarity alone is sufficient.

CHU-50. In contrast, on the more challenging CHU-50 dataset (Table 5), which contains a higher proportion of implicit phenotype mentions, hybrid reranking yields statistically significant improvements over Euclidean retrieval. Paired bootstrap resampling confirms significant gains in weighted recall at Top-1 and at larger candidate depths, as well as consistent and statistically significant improvements in weighted MRR across all evaluated values of k (95% confidence intervals). These results demonstrate that incorporating hierarchical

structure substantially improves both retrieval and ranking of clinically relevant phenotypes in real-world clinical narratives.

D Training Settings

All model training was conducted on a single RTX A3000 GPU, both to accommodate budget constraints and to reduce energy consumption for environmental considerations. Table 6 indicates the training settings for the hyperbolic model training.

Parameter	Value
Number of training epochs	20
Train batch size	32
Eval batch size	64
Learning rate	1e-5
Clustering loss weight	1.0
Clustering loss margin	5.0
Centripetal loss weight	1.0
Centripetal loss margin	0.5
Gradient accumulation steps	8

Table 6: Hyperbolic Training Hyperparameters.

The training settings for the ColbertV2 fine-tuning on HPO is presented in Table 7.

Parameter	Value
Train batch size	8
Learning rate	1e-5
Number of training epochs	2
Max query length	32
Max document length	128
Triplet loss margin	0.3
Gradient accumulation steps	2

Table 7: ColBERTv2 Training Hyperparameters.

E Parameters Settings

For the hierarchy-aware metrics, parameters α (4) and β (5) were set empirically through random assessment by clinicians, who reviewed diverse samples to ensure clinical relevance. We verified that varying them in reasonable ranges does not change the model ranking, and all standard unweighted metrics are also reported. This ensures robustness and prevents metric-induced bias.

Regarding γ parameter (3), we conducted additional ablation studies varying the ratio between Euclidean and hyperbolic components in the hybrid model to assess their respective contributions for the ID-68 and CHU-50 dataset.

As shown in Figures 8 and 9, our findings indicate that except for very small values of γ (which consistently underperform), recall performance

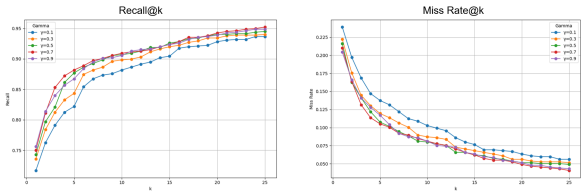


Figure 8: Recall@k and Miss Rate@k for the ID-68 dataset for different values of the weighting parameter γ, which balances Euclidean and hyperbolic distances in the hybrid reranking score.

shows no clear trends across the different γ settings. Among them, $\gamma = 0.5$ and $\gamma = 0.7$ produce the best overall results.

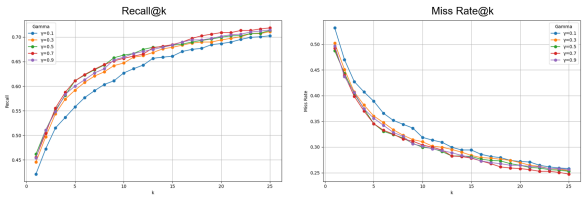


Figure 9: Recall@k and Miss Rate@k for the CHU-50 dataset for different values of the weighting parameter γ, which balances Euclidean and hyperbolic distances in the hybrid reranking score.

Interestingly, for the CHU-50 dataset (characterized by a high proportion of implicit references) the hybrid model with $\gamma = 0.5$ performs identically to, and sometimes better than, $\gamma = 0.7$. This suggests that, in such cases, the hyperbolic component is more effectively leveraged, likely due to its ability to capture hierarchical or indirect relationships.

Given these marginal differences, and to maintain simplicity and interpretability, we fix $\gamma = 0.5$ in the main experiments presented in the paper.

F Relationships Taxonomy

Figure 10 presents the type of relationships considered in the analysis of ontological relationships between the candidates and a given target phenotype (Orange). If the candidate is not in the same branch from the ontology root (*Phenotypic abnormality*), we consider that no path exist between them, as the first level of the ontology refers to very different classes of abnormalities (e.g.: Abnormality of the musculoskeletal system, Abnormality of the nervous system, Abnormality of the cardiovascular system, ...).

While *Ancestor* and *Descendant* are common types, *Cousin* is a broader notion of relationships referring to candidates that have a common ances-

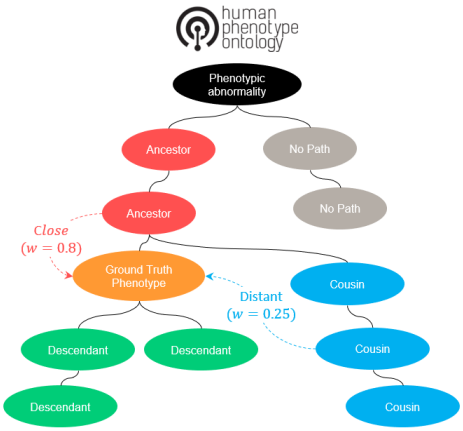


Figure 10: Relationships taxonomy. Given a target phenotype (orange), nodes are categorized as descendants (green), ancestors (red), cousins (blue), or outside the branch (grey). Relative closeness follows the relationship scores defined in Section 5.2.

tor with the target. Finally, a *Close Relationships* type refers to candidates whose relationships score (Equation 4) with the target is above a threshold of 0.7.

G Data Verification

As described in B, training data were generated using ChatGPT-4o-mini to produce sentences from HPO phenotypes with a crafted prompt 18. The generated data were manually evaluated by our clinicians.

We first randomly sampled 50 sentences ensuring coverage across different branches of the HPO ontology. Each sentence was independently reviewed by two clinicians (blind annotation). The annotation guidelines were as follows:

- Meaningfulness: The sentence should be meaningful and coherent.
- Phenotype Reference: The sentence should refer to the input phenotype, either explicitly or implicitly.
- Clinical Realism: The sentence should resemble real-world clinical report formulations.

Annotators selected among three labels: 1 for sentences fully meeting all criteria, 0 if none were met, and 0.5 for partial or ambiguous phenotype references. To assess agreement on this ordinal scale, we used quadratic-weighted Cohen's kappa, yielding a score of 0.64, indicating strong reliability despite the small, imbalanced dataset. Notably, 88% of sentences were rated as high quality. The Cohen's Kappa of 0.64 suggests a strong reliability beyond chance, especially with such small annotation set and label imbalance. The confusion matrix (Figure 11) shows strong agreement on the highest score (1), and most disagreements occur between adjacent categories (e.g., 1 vs. 0.5), suggesting that when annotators differ, they still tend to assign similar quality levels.

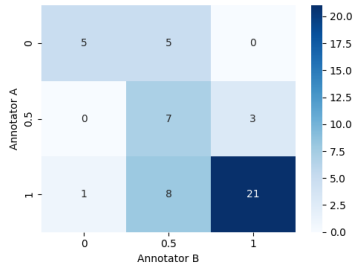


Figure 11: Annotation confusion matrix. Although the class distribution is imbalanced (majority class 1), annotators show strong agreement across classes.

We acknowledge that 50 sentences represent a small sample, but the randomized selection across diverse ontology branches and the high level of agreement observed provide a reasonably robust basis for estimating overall data quality.

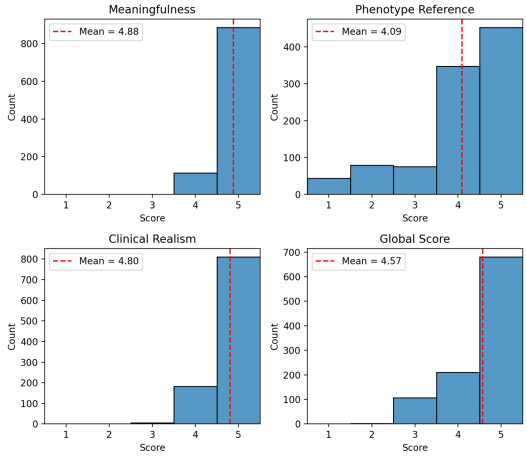


Figure 12: Distribution of LLM-as-a-judge scores across criteria. Scale are from 1 (poor quality) to 5 (excellent quality). Red lines represent the means.

To further support this assessment, we conducted an LLM-as-a-judge evaluation on a larger set of 1000 sentences across the three criteria (Figure 12). While the results should be interpreted with caution, they reinforce the observation that the vast majority of generated sentences are of high quality.

H Additional Results

Cross-Ontology Evaluation

To assess the generalizability of our workflow beyond a single ontology, we conducted additional experiments using a hyperbolic model fine-tuned on the SNOMED ontology instead of HPO. This cross-ontology evaluation tests whether our approach can leverage hierarchical structure from different sources and whether combining ontologies can improve retrieval and ranking. While both ontologies are biomedical, this experiment provides an initial step toward demonstrating the robustness and adaptability of the workflow to diverse hierarchical knowledge bases. Future work will extend this analysis to non-biomedical domains.

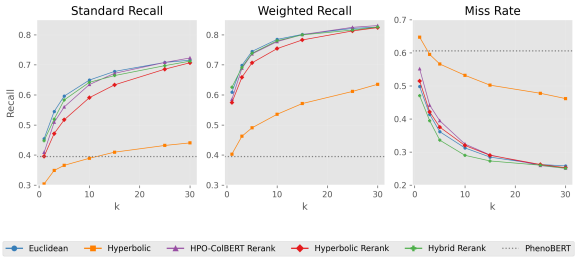


Figure 13: Recall and miss rate on ID-68 using SNOMED-trained hyperbolic embeddings. Exact-match and weighted recall are reported for raw dense retrieval (Euclidean, Hyperbolic) and reranked outputs (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate denotes the proportion of unretrieved ground-truth phenotypes.

As shown in Figure 13, the SNOMED-based hyperbolic model underperforms compared to the HPO-based hyperbolic model, which is expected given that the target phenotypes are defined within the HPO ontology. Interestingly, the hybrid approach exhibits a slight improvement on the ID-68 dataset. This counterintuitive result suggests that, despite the SNOMED model being less domain-specific, it may introduce complementary information that is not captured by the HPO-based hyperbolic models. The combination of semantic similarity and the broader, more general structure encoded by SNOMED embeddings could help differentiate candidates in challenging cases, leading to improved ranking performance.

Conversely, the hybrid model demonstrates reduced accuracy with SNOMED on the CHU-50 dataset. Figure 14 shows recall and miss rate for CHU-50 dataset when using SNOMED hyperbolic model. As expected, the performance is lower than with the HPO hyperbolic model. However, we can note that the miss rate is better with the hybrid approach, suggesting that the hyperbolic model's structure (even when not domain-specialized) may still be effective at ensuring broad coverage of the ontology, due to the hierarchical nature of hyperbolic embeddings.

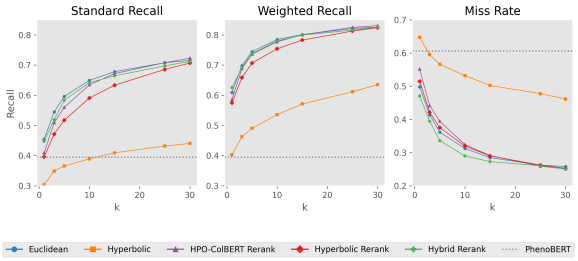


Figure 14: Recall and miss rate on CHU-50 using SNOMED-trained hyperbolic embeddings. Exact-match and weighted recall are reported for raw dense retrieval (Euclidean, Hyperbolic) and reranked outputs (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate denotes the proportion of unretrieved ground-truth phenotypes.

These results highlight the potential for our workflow to integrate information from multiple ontologies, supporting its applicability to a wide range of hierarchical entity linking tasks.

CHU-50 Ontology Metrics

The average number of hops between candidates and target phenotypes is shown in Figure 15. The rank scale is up to 50 so the robustness of the hyperbolic model for higher ranks is highlighted.

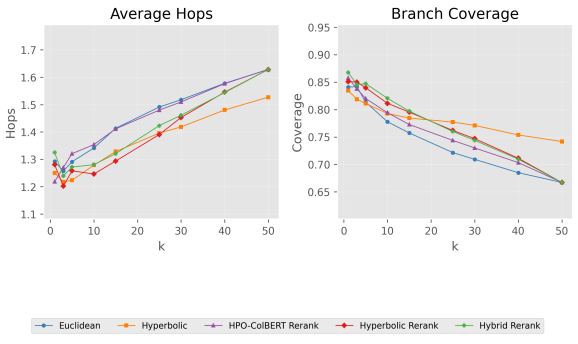


Figure 15: Ontological Structure Metrics on the CHU-50 dataset across *Top-k* candidates. Left: average number of hops between the candidates and the target phenotypes. Right: proportion of candidates within the ontology branch of the target phenotypes.

Finally, the distribution of relationships types for the CHU-50 dataset is shown in Figure 16.

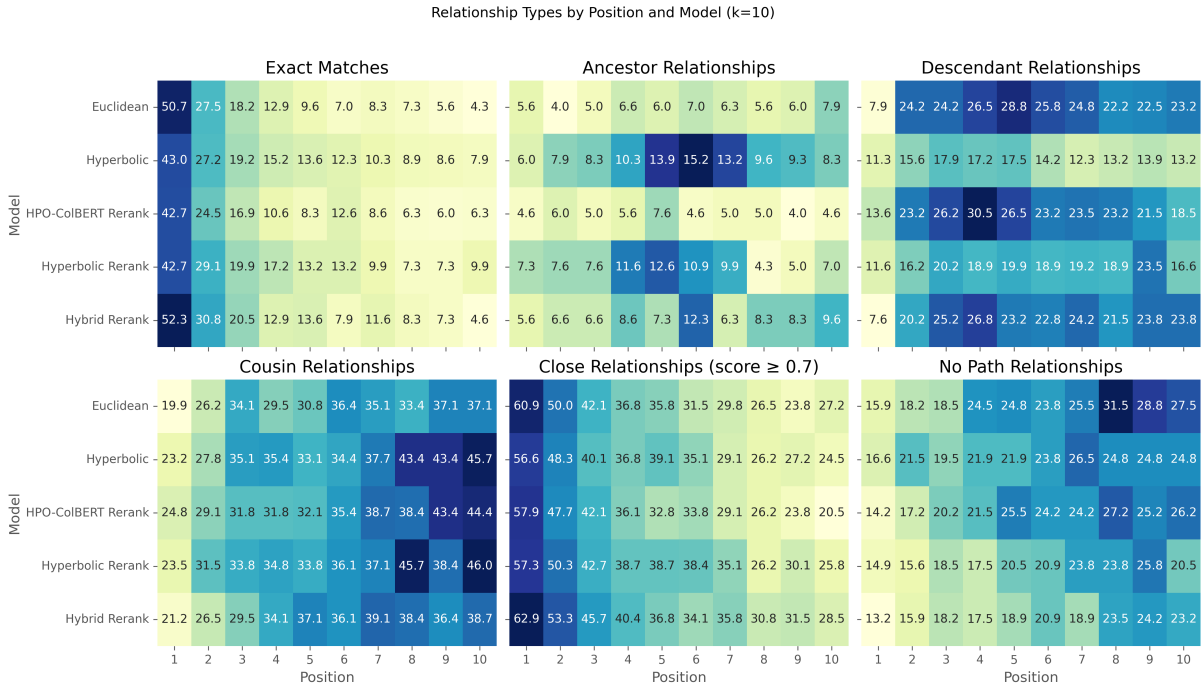


Figure 16: Distribution of relationship types by position and model for Top-10 candidates (k=10) on CHU-50 dataset. The heatmaps show percentages for six relationship categories (Appendix F) between predicted and ground truth phenotypes. For closeness relationships, only those with scores above 0.7 (Equations (4) and (5)) are considered. Rows correspond to different models, and columns represent candidate positions ranked from 1 to 10. Higher values indicate stronger presence of the relationship type at the given rank and model.

I Prompts

The prompt used for span detection step is shown in Fig. 17.

Span Detection

You are an experienced clinician with an exhaustive knowledge of human phenotypes ontology.

Given **{text}**, you must identify the spans related to possible phenotypes, either explicitly or implicitly. You should keep in the span all words related to the phenotype that should be informative (such as negation or adjective).
You may reformulate the span if needed. If you don't detect any span or if you don't know, don't try to make up an answer, just write 'None'.

Figure 17: Full prompt used for span detection (variable in blue).

Late-interaction training data are built through automatic generation of sentences from HPO. This allows to have clinically-relevant sentences for all phenotype labels and synonyms from the ontology. In addition, we crafted prompt so that implicit and measurements references are generated. The corresponding prompt is shown in Fig. 18.

As the goal was to generate spans related to phenotypes, we further query a LLM from the previously generated sentences to get the spans related to the corresponding HPO labels. The prompt used in this step is described in Fig. 19.

Finally, the prompt used for the LLM-as-a-Judge evaluation is presented in Fig. 20.

All prompts were used with the GPT-4o-mini model. The formatting preserves line breaks and spacing as passed to the model.

HPO Sentence Generation

For each HPO label, produce {sentence_count} purely observational sentences, referencing the phenotype explicitly or implicitly.

Requirements

- Use both the main HPO label and all provided synonyms for explicit references.
- At least {implicit_count} sentences must be implicit references (avoid using the label or synonyms).
- Ensure diversity in perspective or detail across sentences (as if from different medical domains and contexts).
- Ensure diversity in sentence openings.
- Use first-person ("I") or neutral ("we") style; do not include any titles (e.g., "Dr. Smith").
- At least two sentences must use passive voice.
- Vary sentence structure, wording and length (some short, some medium, some extended).
- Avoid overuse of "the patient" ($\leq$ 2 uses). Use pronouns ("he," "she," "they") or a fictitious name (first name or "Mr./Mrs." + last name) for diversity.
  - At least 1 sentence must use a fictitious name instead of "the patient" or a pronoun.
- Occasionally include measurements/tests (e.g., mg/dL, $<$ 1st percentile), up to {measures_count} total.
- No interpretive language ("suggesting," "indicative of"); only factual observations.
- No professorial explanations; just present observations.
- Return a bulleted list ("- "). Clinical shorthand is fine.

Context

- **HPO label:** {hpo_label}
- **Definition:** {definition}
- **Synonyms:** {synonyms_str}

Figure 18: Full prompt used for automatic generation of sentences from HPO (variables are shown in blue).

Span Detection from Generated Sentences

Extract the text span that best describes or indicates the phenotype "{phenotype_data['hpo_label']}" from this sentence.

Requirements

1. Capture the complete clinical observation or symptom
2. Include relevant context that helps understand the phenotype
3. Be specific to the actual medical condition
4. Be concise while maintaining clinical accuracy
5. Exclude patient names, temporal markers, or examination context
6. Focus on the actual phenotypic finding

Sentence: "{phenotype_data['sentence']}"
Extract only the relevant span without any additional commentary.

Figure 19: Full prompt used for span detection from generated sentences (variables are shown in blue). The object *phenotype_data* refers to a dataframe with the original HPO phenotype (*hpo_label*) and the corresponding generated sentence (*sentence*).

LLM-as-a-Judge Evaluation

You are an expert clinical language model evaluating sentences linked to HPO terms. Evaluate the sentence below using these criteria. Each score must be an integer from 1 to 5 (5 = excellent, 1 = poor):

1. **Meaningfulness** – Is the sentence meaningful and coherent?
2. **Phenotype Reference** – Does the sentence refer to the phenotype "**{hpo_label}**" (**{hpo_id}**), explicitly or implicitly?
3. **Clinical Realism** – Could this sentence appear in a real clinical report?

Then provide a **global score** (also 1–5) summarizing the overall quality of the sentence with respect to all three criteria.

Respond ONLY in this JSON format, filling in each `<score>` with your evaluation:

```
{
  "meaningfulness": <score>,
  "phenotype_reference": <score>,
  "clinical_realism": <score>,
  "global_score": <score>,
  "comment": "Brief explanation"
}
```

Sentence: "**{sentence}**"

Figure 20: Full prompt used for LLM-as-a-Judge evaluation.

Agentic Feature Selection via LLM for Epileptic Seizure Detection

Aizierjiang Aiersilan ✉ **Xiaodong Qu**
The George Washington University
alexandera@gwu.edu

Abstract

Automated epileptic seizure detection from electroencephalography (EEG) signals is a clinically important task in which feature selection is typically performed using purely statistical criteria. We investigate whether a small instruction-tuned large language model (LLM) can guide iterative feature selection for binary seizure detection on the Epileptic Seizure Recognition dataset (11,500 samples, 178 features). The LLM agent (Qwen2.5-1.5B-Instruct) receives five complementary statistical summaries and selects a feature subset through multi-round reasoning. The agent achieves 96.5% accuracy and 0.911 F1 with 40 features, compared to 97.9% accuracy and 0.946 F1 for the best full-feature baseline (SVM-RBF on 178 features). Critically, 39 of the agent's 40 features coincide with the top-39 mutual-information features, and a deterministic Top-39 MI filter, evaluated by the same Random Forest classifier, attains the same 96.5% accuracy and 0.911 F1. We therefore present this work as an empirical baseline: at the 1.5B-parameter scale, the LLM behaves close to a univariate MI ranker. We situate the result against the recent LLM-based feature selection literature and enumerate the ablations and multi-dataset extensions required to determine whether larger or domain-specialized LLMs add value beyond statistical filtering.

Code: github.com/Ezharjan/AgenticFS4EEG.

1 Introduction

Epilepsy affects approximately 50 million people worldwide and is characterized by recurrent, unprovoked seizures (Shoeb, 2009). Electroencephalography (EEG) is the primary non-invasive diagnostic modality, but manual review of long-term recordings is slow and subject to inter-rater variability. Automated detectors based on classical machine learning such as random forests (Breiman, 2001), support vector machines (Cortes and Vapnik, 1995), and gradient boosting (Friedman, 2001) achieve strong performance on benchmark EEG tasks (Shimanto, 2017), but operating on the full feature space retains redundant or noisy dimensions, limiting interpretability and efficiency. Feature selection is therefore a critical preprocessing step (Guyon and Elisseeff, 2003), and is traditionally performed using statistical criteria such as mutual information (MI) (Kraskov et al., 2004), ANOVA F-scores, or wrapper-based search, none of which carry semantic understanding of the features or the domain. Recent LLM advances open the possibility of incorporating natural-language reasoning into such pipelines (Zhao et al., 2025; Wang et al., 2026).

We propose an *agentic LLM-guided feature selection* framework for binary epileptic seizure detection. The LLM agent receives five complementary statistical measures for each feature and produces a feature subset through five rounds of iterative reasoning. We evaluate the agent against three full-feature baselines and against a purely MI-driven filter baseline. Our contributions are: (1) an agentic framework coupling LLM reasoning with five complementary statistics for iterative EEG feature selection (Algorithm 1, Appendix D); (2) an evaluation showing that compressing the feature space from 178 to 40 features (a 77.5% reduction) via LLM mediation incurs only a 1.4 percentage-point accuracy gap relative to the best full-feature baseline; and (3) a controlled comparison against a deterministic Top-39 MI filter, evaluated with the same Random Forest classifier on the same split, showing that at the 1.5B-parameter scale the LLM agent's selection coincides with the MI ranking on 39 of 40 features and produces statistically indistinguishable downstream metrics. We present this as an empirical baseline rather than as evidence that LLMs add value beyond a univariate statistical filter on this benchmark, and enumerate the ablations and extensions required to revisit the question in Section 8.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 64–74
July 3-4, 2026 ©2026 Association for Computational Linguistics

2 Related Work

2.1 EEG-Based Seizure Detection

The Bonn University EEG database (Andrzejak et al., 2001) and its derivative Epileptic Seizure Recognition dataset (Shimanto, 2017) are primary benchmarks for seizure detection. Classical approaches extract time-, frequency-, and wavelet-domain features and apply supervised classifiers (Hastie, 2009), with random forests (Breiman, 2001) and nonlinear-kernel SVMs (Cortes and Vapnik, 1995) consistently achieving top performance. More recently, CNNs and transformers have been applied to raw EEG signals (Yang and Modesitt, 2023; Konkar and Qu, 2026), but typically require larger datasets and offer less interpretability than feature-based methods. Our work retains the feature-based paradigm and focuses on *how* features are selected rather than how they are extracted.

2.2 Feature Selection for EEG

Feature selection methods for EEG span filter, wrapper, and embedded approaches (Guyon and Elisseeff, 2003). NeuroFeat (Choudhury et al., 2026) adaptively selects features based on signal characteristics. Presacan et al. (Presacan et al., 2025) review explainable AI methods for EEG, and Lai et al. (Lai et al., 2025) note that feature engineering remains essential for small-to-medium datasets where deep models tend to overfit. These works rely on statistical or model-intrinsic criteria; we instead delegate selection to an LLM that can jointly interpret multiple statistical summaries.

2.3 LLM Agents for Scientific Data Analysis

EEGAgent (Zhao et al., 2025) and NeuroWeaver (Wang et al., 2026) are recent end-to-end LLM-driven frameworks for EEG analysis, and Murungi et al. (Murungi et al., 2023) survey LLM-assisted EEG interpretation. To our knowledge, neither EEGAgent nor NeuroWeaver reports detection metrics on the specific Epileptic Seizure Recognition benchmark used here, so direct numerical comparison is not currently possible. Our work differs by focusing on feature selection rather than automating the whole pipeline.

2.4 LLM-Based Feature Selection

LLM-Select (Jeong et al., 2024) queries an LLM with feature names and short descriptions to rank features zero-shot. CAAFE (Hollmann et al., 2023) uses an LLM to iteratively generate and evaluate new feature transformations with downstream classifier feedback. LLM-FS-Agent (Bal-Ghaoui and Sabri, 2025) adopts a deliberative multi-role LLM architecture in which several agents debate each feature's inclusion. We share the paradigm of delegating selection to an LLM but differ in three ways: (i) we operate on raw amplitude time-step features (X1–X178) without semantically meaningful names, so zero-shot name-based ranking is not applicable and the agent must reason over numeric statistics; (ii) unlike CAAFE we do not synthesize features but only select from a fixed set, giving a stricter test of the LLM's marginal contribution beyond a univariate ranker; and (iii) in contrast to LLM-FS-Agent's multi-agent debate, we use a single 1.5B-parameter instruction-tuned model with a structured statistical prompt. Our findings (Section 6) thus provide an empirical baseline complementary to the broadly positive results reported by larger or more elaborate LLM-FS systems.

3 Dataset

We use the Epileptic Seizure Recognition dataset (Shimanto, 2017), a publicly available benchmark derived from the Bonn University EEG database (Andrzejak et al., 2001). It contains 11,500 EEG segments, each represented by 178 time-domain amplitude features (sampled at 173.61 Hz) and a class label $y \in \{1, 2, 3, 4, 5\}$. The five classes encode recording conditions: y=1 denotes ictal seizure activity; y=2 and y=3 correspond to seizure-free recordings from the epileptogenic zone and hippocampal formation, respectively; and y=4 and y=5 correspond to surface EEG from healthy subjects. Each class contains exactly 2,300 samples.

Following the standard experimental protocol (Shimanto, 2017), the task is converted to binary classification: class 1 (seizure, n = 2,300) vs. classes 2–5 (non-seizure, n = 9,200), yielding a moderate 4:1 class imbalance ratio.

4 Methodology

The proposed pipeline comprises five stages: preprocessing, baseline training, LLM-guided feature selection, evaluation, and visualization. Figure 2 illustrates the overall architecture, and Figure 1 details the iterative reasoning loop at the core of the agent.

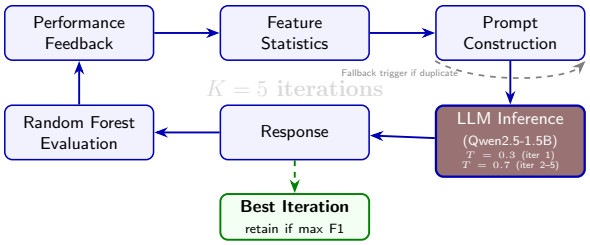


Figure 1: LLM Agent iterative reasoning loop. At each iteration, statistical summaries are compiled into a structured prompt, the LLM proposes a feature subset, a Random Forest evaluates the subset, and performance feedback informs the next round. The iteration with the highest F1-score is retained. A diversity fallback (dashed self-loop) activates if any proposal duplicates a prior selection.

4.1 Preprocessing

The 178 amplitude features (X1–X178) form the feature matrix $\mathbf{X} \in \mathbb{R}^{11500\times178}$. Labels are binarized such that $y{=}1$ maps to the positive (seizure) class and $y \in \{2,3,4,5\}$ to the negative (non-seizure) class. All features are standardized to zero mean and unit variance, with statistics computed on the training set and applied to the test set. A stratified 80/20 split preserves class proportions across training ($n{=}9{,}200$) and test ($n{=}2{,}300$) partitions.

4.2 Baseline Models

Three classifiers are trained on the full 178-dimensional standardized feature set as performance upper bounds, all implemented with scikit-learn (Pedregosa et al., 2011): a Random Forest (RF) ensemble of 200 trees with depth 20 and $\sqrt{d}$ features per split (Breiman, 2001); an SVM with RBF kernel ($C{=}10$, $\gamma{=}$scale) (Cortes and Vapnik, 1995); and a Gradient Boosting (GB) ensemble of 200 shallow trees (depth 5, learning rate 0.1) (Friedman, 2001).

4.3 LLM-Guided Feature Selection Agent

The agent is the methodological focus of this work. It is an iterative reasoning loop that combines information-theoretic feature profiling with LLM decision making to produce a compact subset for downstream classification; Algorithm 1 formalizes the procedure.

4.3.1 Feature Statistics Computation

For each of the 178 features x_j, five complementary statistics are computed on the training set. Mutual information (MI) is estimated with the k-nearest-neighbor method of Kraskov et al. (Kraskov et al., 2004) and captures nonlinear feature-label dependency. Pearson correlation captures the linear component of the same association, allowing the agent to distinguish purely linear from nonlinear discriminative structure. Variance is included to confirm the absence of degenerate (near-constant) features after standardization. Random Forest importance, the mean decrease in impurity from a preliminary RF (Breiman, 2001), complements the filter statistics by reflecting utility within an ensemble classifier. The ANOVA F-score (between-class over within-class variance) signals strong univariate class separation. These five measures span information-theoretic (MI), linear-statistical (correlation, F-score, variance), and model-based (RF importance) views, so the LLM has complementary evidence to cross-reference rather than a single ranking to imitate. Features are presented to the LLM sorted by MI in descending order.

4.3.2 Prompt Architecture

At each iteration k, the agent constructs a structured prompt with four components. A system message establishes the LLM's role as a feature selection expert and constrains outputs to a JSON list of feature names. A statistical summary table lists, for the top-40 features by MI, each feature's MI score, Pearson correlation, RF importance, and F-score. For iterations $k \geq 2$, the prompt is augmented with performance feedback (the prior iteration's accuracy, F1, AUC; the five most and five least important features by RF importance; and the full previous feature list), and with a mutation instruction directing the model to drop at least five low-importance features and replace them with high-MI or high-F-score alternatives, returning exactly 30 features as a JSON list. This structure mimics a human analyst's workflow: inspect, then refine based on downstream performance. Exact prompt templates are in Appendix D.

4.3.3 Temperature Schedule and Diversity Mechanisms

Exploration is controlled via a two-phase temperature schedule. Iteration 1 uses $T{=}0.3$ to produce a stable, high-confidence baseline selection, and iterations 2–5 use $T{=}0.7$ to encourage exploration. A diversity safeguard ensures no two iterations produce identical subsets: an exact match triggers a deterministic mutation that drops the 5 lowest-MI features and adds the highest-MI features not yet selected. The LLM's output is also clamped to the

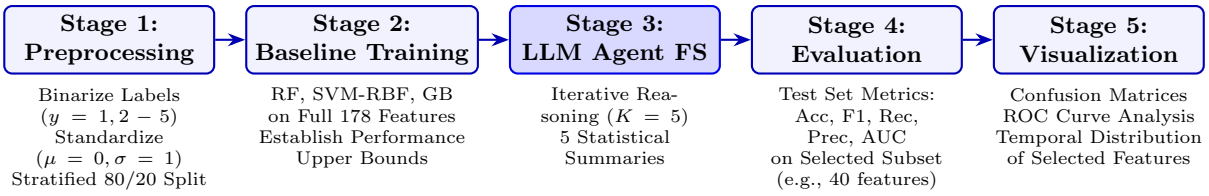


Figure 2: Five-stage experimental pipeline. The LLM Agent Feature Selection stage (emphasized) is the methodological focus of this work.

loose size range $[20, 80]$ (small subsets are padded with high-MI features; large ones are truncated).

Note on requested vs. returned size. The prompt requests *exactly* 30 features, but the post-processor only enforces $[20, 80]$ and otherwise preserves the model's choice. Because Qwen2.5-1.5B-Instruct (Qwen Team, 2025) does not strictly satisfy length constraints at this scale, the returned subsets ranged from 20 to 40 features across iterations (Table 2); we deliberately did not pad or truncate to 30, since doing so would overwrite the model behavior we wished to study. The discrepancy between the prompted 30 and the retained 40 (iteration 1) is therefore a faithfully reported instruction-following limitation, discussed further in Section 7.

4.3.4 Iterative Refinement Loop

The agent executes $K{=}5$ iterations (Algorithm 1). At each iteration the LLM receives the statistics (and, from $k{\geq}2$, the prior iteration's feedback), proposes a subset, which a Random Forest (200 trees, depth 20) trains and scores on accuracy, F1, and AUC; the highest-F1 iteration is retained. If the LLM fails to return a valid response, a statistical fallback selects the top-k features from one of the five precomputed statistics, cycling through MI, F-score, RF importance, correlation magnitude, and variance across successive fallbacks.

Algorithm 1 Agentic LLM-Guided Feature Selection

Require: Matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, labels $\mathbf{y}$, LLM $\mathcal{M}$, iterations $K = 5$
Ensure: Best feature subset S^*

Compute stats $\mathbf{s}_j = (\text{MI}_j, \rho_j, \sigma_j^2, \text{RF}_j, F_j)$ and rank by MI
$S^* \leftarrow \varnothing, F_1^* \leftarrow 0, \mathcal{H} \leftarrow \varnothing$ ▷ Init best set, score, and history
for $k = 1$ **to** K **do**
  $T_k \leftarrow (k = 1)?0.3 : 0.7$ ▷ Temperature schedule
  $p_k \leftarrow \textsc{BuildPrompt}(\mathbf{s}, \mathcal{H}, \text{fb}_{k-1})$
  $r_k \leftarrow \mathcal{M}.\text{gen}(p_k, T_k)$ ▷ LLM sampling
  $S_k \leftarrow \textsc{ParseAndClamp}(r_k, 20, 80)$ ▷ Enforce size constraints
  if $S_k \in \mathcal{H}$ **then** $S_k \leftarrow \textsc{MutateFallback}(S_k, \mathbf{s})$ ▷ Handle stagnation
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{S_k\}$
  $(\text{acc}_k, \text{F1}_k, \text{auc}_k, \mathcal{I}_k) \leftarrow \textsc{Evaluate}(S_k, \mathbf{X}, \mathbf{y})$
  if $\text{F1}_k > F_1^*$ **then**
    $S^* \leftarrow S_k, F_1^* \leftarrow \text{F1}_k$ ▷ Update global optimum
  $\text{fb}_k \leftarrow (\text{acc}_k, \text{F1}_k, \text{auc}_k, \mathcal{I}_k)$ ▷ Feedback for next iteration
return S^*

4.3.5 Evaluation of Selected Features

The final feature subset trains a Random Forest classifier (200 trees, depth 20, identical to the full-feature RF baseline) on the training set and is evaluated on the held-out test set; the LLM agent uses RF as its internal evaluator, so final reported metrics are computed on unseen data.

Top-k MI Filter baseline. To isolate the LLM's marginal contribution, we evaluate a deterministic Top-k MI Filter baseline that selects the k features with highest training-set MI with the label, with $k{=}39$ chosen to closely match the agent's 40-feature budget. For full comparability, this MI-filter subset is evaluated using *exactly the same Random Forest classifier* (200 trees, depth 20, seed 42) used to score the agent's selection; only the input feature columns differ. Any gap between the two is therefore attributable to feature choice alone.

5 Experimental Setup

We use Qwen2.5-1.5B-Instruct (Qwen Team, 2025) as the LLM backbone via a hosted inference endpoint with a 500-token output limit. For each method we report accuracy, precision, recall (sensitivity), F1, and AUC; confusion matrices are pro-

vided for all methods. Experiments ran on an NVIDIA RTX 5060 GPU with Python 3.x and scikit-learn (Pedregosa et al., 2011), with random seed 42 for reproducibility.

Choice of LLM backbone. We deliberately use a small (1.5B), general-purpose instruction-tuned model rather than a larger or biomedical LLM. First, small open-weight models are realistic for clinical edge environments where data residency and latency constraints discourage routing EEG statistics to large hosted models. Second, our goal is not peak accuracy but characterizing what an off-the-shelf lightweight LLM contributes *above* a purely statistical filter; a small contribution at this scale is itself an informative empirical baseline for the LLM-FS literature (Section 2.4). Third, available medical LLMs at comparable scale are predominantly clinical-text models tuned for narrative question answering rather than reasoning over numeric tables of unnamed time-step features, so their advantage here is not obvious in advance. Scaling and domain-specialized backbones are listed as future work in Section 8.

6 Results

6.1 Overall Performance Comparison

Table 1 presents the classification performance of all methods on the held-out test set (2,300 samples).

Table 1: Binary seizure detection performance on the test set. Full-feature baselines use all 178 features; the Top-39 MI Filter uses 39 features and the LLM agent uses 40 features.

Method	Acc ↑	Prec ↑	Rec ↑	F1 ↑	AUC ↑
Random Forest	0.971	0.960	0.891	0.925	0.996
SVM-RBF	**0.979**	0.956	**0.937**	**0.946**	0.996
Grad. Boosting	0.974	**0.967**	0.902	0.934	**0.996**
Top-39 MI Filter (39 ft.)	0.965	0.922	0.900	0.911	0.993
LLM Agent (40 ft.)	0.965	0.924	0.898	0.911	0.993

All methods exceed 96.4% accuracy and 0.992 AUC, confirming the dataset's high discriminability. SVM-RBF achieves the highest accuracy (97.9%), F1 (0.946), and recall (93.7%, the highest among all methods), missing the fewest seizure events. Gradient Boosting achieves the highest precision (96.7%) and AUC (0.996), indicating low false-positive rate at the cost of slightly lower recall. Random Forest balances precision (96.0%) and recall (89.1%), with competitive AUC (0.996). The LLM Agent isolates 40 features, dropping roughly three-quarters of the input space, and still preserves strong predictive power with only a minor metric drop.

The Top-39 MI Filter, evaluated with the identical Random Forest classifier, attains comparable results from 39 features selected by mutual information alone. The LLM's chosen set is a strict superset of these 39 features, adding only X61; the overlap is therefore 39/40 (97.5%). Across Table 1's metrics the two methods differ by at most 0.002 on precision and recall and are identical to three decimals on accuracy, F1, and AUC. The LLM's marginal contribution beyond a univariate MI ranking at this scale is thus not statistically meaningful on this benchmark; we treat this as the central empirical finding and discuss its implications in Section 7.

Clinically, the LLM agent's 89.8% recall is below the full-feature baselines, reflecting the cost of dimensionality reduction; acceptability is deployment-specific (screening tolerates lower recall, ICU monitoring demands maximal sensitivity, see Section 7.4).

6.2 Confusion Matrix Analysis

Figure 3 presents the confusion matrices for all methods.

SVM-RBF has 29 false negatives (FN) and 20 false positives (FP), the best sensitivity. The LLM agent has 47 FN and 34 FP, a modest increase but a 10.2% false-negative rate on the seizure class. The Top-39 MI Filter has 46 FN and 35 FP, comparable to the agent. Gradient Boosting has the fewest FP (14) but 45 FN, consistent with its high-precision profile.

6.3 ROC Curve Analysis

Figure 4 shows the receiver operating characteristic (ROC) curves for all methods.

All five ROC curves are nearly indistinguishable, hugging the upper-left corner. AUC ranges from 0.993 (Top-39 MI Filter) to 0.996 (Gradient Boosting), a 0.004 spread, so the reduced-feature methods preserve nearly all ranking quality of the full-feature models.

6.4 Feature Selection Analysis

The LLM agent selected 40 features from the 178 available. Figure 5 shows the RF feature importances across all 178 features, providing context for the agent's selection.

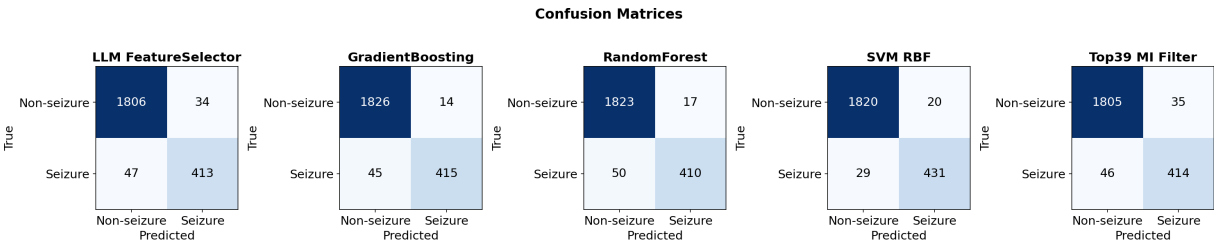


Figure 3: Confusion matrices on the test set for all five methods.

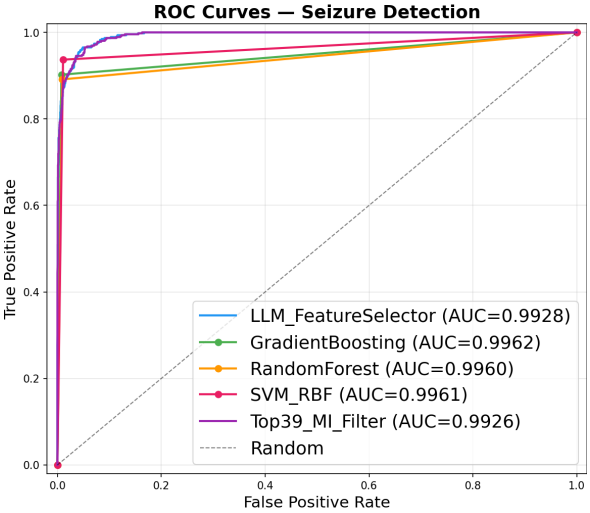


Figure 4: All methods' ROC curves achieve AUC $>$ 0.992.

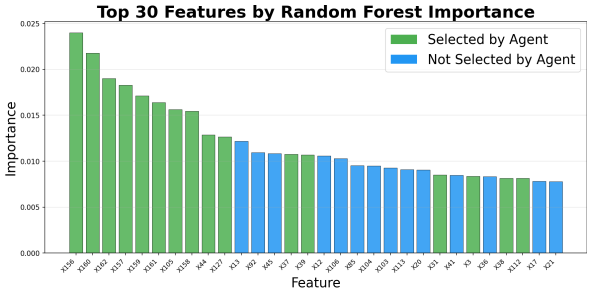


Figure 5: Random Forest feature importances (mean decrease in impurity) for all 178 features.

The five highest-MI features in the LLM's selection are X44, X172, X2, X160, and X39, all with MI above 0.199. The 40 selected features span the full EEG segment (X2 to X174), indicating the LLM captured signal dynamics across the whole recording window rather than concentrating on one region. Figure 6 visualizes this spread.

As noted earlier, the agent's chosen 40-feature subset perfectly encompasses the Top-39 MI Filter selections, supplementing them with only X61. This near-total alignment is explored structurally in Section 7.

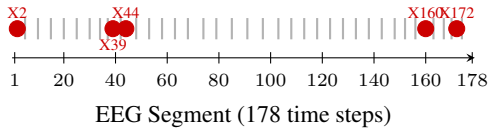


Figure 6: Temporal distribution of the 40 LLM-selected features across the 178-step EEG segment. Red circles mark the five highest-MI features; gray ticks mark the remaining selected features. The selection spans the full recording window.

6.5 Agent Iteration Behavior

Figure 7 and Table 2 present the full performance profile and subset sizes of the LLM agent across the 5 iterations.

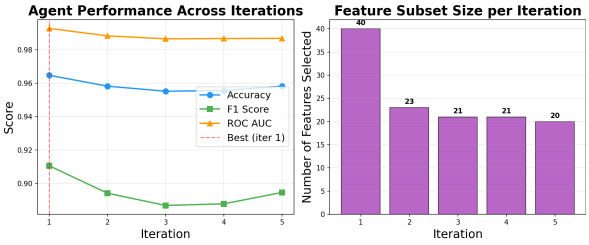


Figure 7: Full performance profile of the LLM agent across 5 iterations.

Table 2: Per-iteration agent outputs. All evaluation metrics are listed for each iteration.

Iter.	T	#Feat. ↓	Acc ↑	F1 ↑	AUC ↑
1	0.3	40	**0.965**	**0.911**	**0.993**
2	0.7	23	0.958	0.894	0.988
3	0.7	21	0.955	0.887	0.987
4	0.7	21	0.956	0.888	0.987
5	0.7	**20**	0.958	0.895	0.987

The temperature schedule and feature-level feedback produced genuinely different subsets across iterations, with no diversity fallback triggered. The best selection was iteration 1 (T=0.3, F1 0.911, 40 features); higher-temperature iterations (T=0.7) selected 20–23 features and underperformed by an F1 range of 0.024 (0.887–0.911). Both regimes

deviated from the prompted 30-feature target, indicating loose length-constraint adherence at this scale. Iteration 5 reached F1 0.895 with only 20 features, slightly above iterations 3–4 (21 features), so subset composition affects outcomes beyond raw count.

6.6 Performance Comparison Overview

Figure 8 compares all methods across the five metrics.

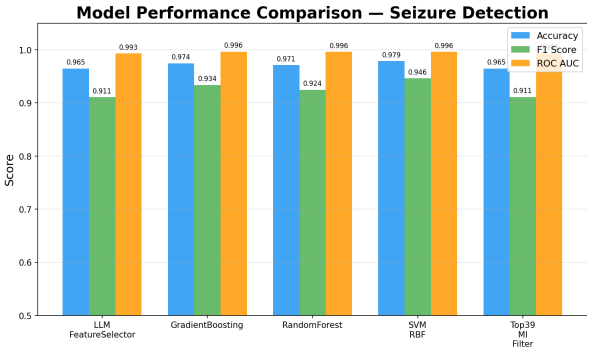


Figure 8: Bar chart comparing Accuracy, F1, and AUC across all five methods.

All methods are tightly clustered on AUC and accuracy; the LLM agent and Top-39 MI Filter perform nearly identically, both slightly below the full-feature baselines, with the gap most visible on precision and recall.

7 Discussion

7.1 Accuracy–Interpretability Trade-off

The LLM-driven pipeline yields competitive detection metrics with a fraction of the features, but this must be qualified: the same compact subset is achievable here by a statistical MI filter, so the interpretability gain reported is best attributed to dimensionality reduction rather than to the LLM. Within that caveat, a small feature set is practically useful in settings where model decisions must be auditable, since a clinician can inspect a concise list of waveform amplitudes against known electrophysiological signatures, whereas auditing all 178 features is impractical at run time. The marginal accuracy drop may be acceptable in screening or outpatient contexts, but is harder to justify in intensive-care monitoring where false-negative costs dominate (Section 7.4).

7.2 LLM Reasoning Capabilities and Limitations

Qwen2.5-1.5B-Instruct (Qwen Team, 2025) reliably parses statistical tables, ranks features by MI and F-score, and emits well-formatted JSON, showing a small LLM can serve as a functional component given a clear, structured prompt. However, four behavioral observations constrain its practical value at this scale (each consolidated in the Limitations of Section 8). First, instruction adherence is weak: the model returned 40 features at low temperature and 20–23 at higher temperatures despite an explicit 30-feature request. Second, the near-complete overlap (39/40, 97.5%) with the top-39 MI features is consistent with the model disproportionately weighting MI, which is presented both as a tabulated value and as the row sort key. This interpretation is only suggested by the overlap; without the no-MI or shuffled-MI ablations in Section 8 we cannot rule out that the LLM rediscovered the MI ranking from the other three columns. Third, iteration did not improve the initial selection (best F1 stays at iteration 1) but did not degrade it either, with F1 spanning only 0.024; we call this robust but non-improving. Fourth, the model produced no chain-of-thought rationale (Wei et al., 2022), because the prompt restricts the output to a JSON list, which is a prompt-design choice rather than a backbone limitation.

7.3 Comparing Statistical Feature Selection

The Top-39 MI Filter, evaluated with the identical Random Forest and split, matched the agent's headline metrics to three decimals on accuracy, F1, and AUC, with at most a 0.002 difference on precision and recall. The equivalence derives directly from the model's selections (39 of 40 features are the top-39 MI features). Because the prompt foregrounds MI both positionally and informationally, the subset can plausibly be reproduced by a copy-the-top-ranked-rows heuristic. Temperature variation does not control for this confound, since temperature affects stochasticity rather than which input cues the model attends to. The cleanest controls (deterministic decoding $T{=}0$, a no-MI or shuffled-MI prompt, and per-iteration quantification of deviation from the rolling top-k MI list) are enumerated as required follow-ups in Section 8. The equivalence still has value as an empirical upper bound on purely MI-driven filtration at a 40-feature budget, and the agentic framework retains structural

advantages, since it can readily incorporate natural-language instructions (e.g., "prefer early temporal features") and be paired with more capable LLMs without architectural changes.

7.4 Comparing Deep Learning Baselines

Recent end-to-end deep models trained on the raw EEG of this benchmark report 98–99%+ accuracy; for example, Guhdar et al. (Guhdar et al., 2025) report a 1D-CNN with multi-head attention surpassing 99%. Our best classical baseline (SVM-RBF, 97.9%) and our LLM-guided 40-feature subset (96.5%) sit below this deep-learning state of the art. We do not claim parity. Instead, the contribution is a small, explicitly enumerated, human-auditable feature subset that exposes which time-step amplitudes drive the decision. End-to-end CNN/transformer detectors offer higher accuracy and recall but operate on hundreds of millions of opaque weights over the raw waveform, which complicates the post-hoc audit our pipeline targets. This trade-off is unfavorable in high-acuity settings (e.g., ICU monitoring); we revisit this in the limitations.

7.5 Claim, Caveat, and Path Forward

Interpretable rationales are arguably the strongest motivation for an LLM in a feature-selection loop, and we acknowledge that this paper's interpretability claim rests on subset size rather than on model-generated narrative. The current prompt (Appendix D) restricts the output to a JSON list, which is a prompt-design choice, not a backbone limitation. A revised prompt requiring per-feature justifications, or an explicit chain-of-thought trace (Wei et al., 2022) preceding the JSON list, would yield qualitative material amenable to neurologist inspection. We have not run that experiment here, and we list eliciting and evaluating such rationales as a primary direction for future work in Section 8.

8 Conclusion

We presented an agentic feature-selection framework for EEG seizure detection in which a small instruction-tuned LLM (Qwen2.5-1.5B-Instruct) iteratively reduces a 178-dimensional input over five statistical measures. On this benchmark the agent attained 96.5% accuracy and 0.911 F1 with 40 features (a 77.5% reduction), versus 97.9% accuracy and 0.946 F1 for the strongest full-feature baseline (SVM-RBF). The headline finding is a negative one: 39 of the agent's 40 features are exactly the top-39 MI features, and a deterministic Top-39 MI filter evaluated by the same Random Forest classifier matches the agent's downstream metrics. At the 1.5B-parameter scale, with MI presented as both a tabulated value and the row sort key, the model behaves close to a copy-the-top-ranked-rows heuristic. We frame this as an empirical baseline that future, more capable LLM-FS systems must beat to justify their added complexity.

Limitations. We list the following limitations so the contribution is not overstated. (L1) MI is presented both as a numeric column and as the row sort key, and we did not run an ablation withholding or shuffling MI, so the agent's behavior cannot be cleanly separated from a univariate MI ranker. (L2) All results are on one saturated benchmark where multiple methods already exceed 96% accuracy, so generalization to multi-channel, longer, or noisier recordings is unsupported here. (L3) Selected features are evaluated only with Random Forest, leaving cross-classifier transfer (notably to SVM-RBF) untested. (L4) The JSON-only output precludes per-feature rationales, so the interpretability claim rests on subset size rather than model narrative. (L5) The model did not respect the prompted "exactly 30 features" constraint (returning 20–40 across iterations). (L6) Recent end-to-end deep models report 98–99%+ accuracy on the same dataset (Guhdar et al., 2025); the recall drop is clinically non-trivial. (L7) Larger LLMs and domain-specialized medical LLMs were not evaluated. (L8) Headline numbers use a single fixed seed, with no variance reported across seeds.

Future Work. The most informative follow-ups are: (1) an MI ablation that removes or shuffles the MI column and uses deterministic decoding ($T{=}0$); (2) per-iteration quantification of how often the LLM's selection deviates from the rolling top-k MI list; (3) cross-classifier transfer of the agent's 40-feature subset to SVM-RBF and Gradient Boosting, with the same reported for the Top-k MI baseline; (4) multi-dataset generalization on CHB-MIT (Guttag, 2010), Siena (Detti, 2020), and EPILEPSIAE (Ihle et al., 2012); (5) scaling to 7B–70B LLMs and to domain-specialized medical LLMs to test for genuine cross-statistic synthesis; (6) prompts that elicit per-feature rationales or full chain-of-thought (Wei et al., 2022) for clinically auditable justifications; (7) recall-targeted refinement that feeds false-negative patterns back into the next iteration; and (8) multi-seed reporting with confidence intervals.

References

Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. 2001. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.

Mohamed Bal-Ghaoui and Fayssal Sabri. 2025. Llm-fs-agent: a deliberative role-based large language model architecture for transparent feature selection. *arXiv preprint arXiv:2510.05935*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Nitin Choudhury, Daisy Das, Deepjyoti Deka, Rajdeep Ghosh, Nabamita Deb, and Ebrahim Ghaderpour. 2026. Neurofeat: An adaptive neurological eeg feature engineering approach for improved classification of major depressive disorder. *Biomedical Signal Processing and Control*, 113:109031.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Paolo Detti. 2020. Siena Scalp EEG Database. *PhysioNet*. Version 1.0.0.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Mohammed Guhdar, Ramadhan J Mstafa, and Abdulhakeem O Mohammed. 2025. Hybrid deep learning model for epileptic seizure classification by using 1d-cnn with multi-head attention mechanism. *arXiv preprint arXiv:2501.10342*.

John Guttag. 2010. CHB-MIT Scalp EEG Database. *PhysioNet*. Version 1.0.0.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Trevor Hastie. 2009. The elements of statistical learning: data mining, inference, and prediction.

Noah Hollmann, Samuel Müller, and Frank Hutter. 2023. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems*, 36:44753–44775.

Matthias Ihle, Hinnerk Feldwisch-Drentrup, César A Teixeira, Adrien Witon, Björn Schelter, Jens Timmer, and Andreas Schulze-Bonhage. 2012. Epilepsiae–a european epilepsy database. *Computer methods and programs in biomedicine*, 106(3):127–138.

Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. 2024. Llm-select: Feature selection with large language models. *arXiv preprint arXiv:2407.02694*.

Aniket Konkar and Xiaodong Qu. 2026. A review of transformer-based and hybrid deep learning approaches for eeg analysis. In *International Conference on Human-Computer Interaction*, pages 391–404. Springer.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138.

Junhong Lai, Jiyu Wei, Lin Yao, and Yueming Wang. 2025. A simple review of eeg foundation models: Datasets, advancements and future perspectives. *arXiv preprint arXiv:2504.20069*.

Nathan Koome Murungi, Michael Vinh Pham, Xufeng Caesar Dai, and Xiaodong Qu. 2023. Empowering computer science students in electroencephalography (eeg) analysis: A review of machine learning algorithms for eeg datasets. In *The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Oriana Presacan, Jaya Ojha, Anis Yazidi, Eric Monteiro, and Pedro G Lind. 2025. A comprehensive review of explainable ai in deep learning algorithms for eeg analysis. *ACM Transactions on Computing for Healthcare*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Harun Shimanto. 2017. Epileptic seizure recognition. https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition. Accessed: 2026-02-18.

Ali Hossam Shoeb. 2009. *Application of machine learning to epileptic seizure onset detection and treatment*. Ph.D. thesis, Massachusetts Institute of Technology.

Guoan Wang, Shihao Yang, Jun-En Ding, Hao Zhu, and Feng Liu. 2026. Neuroweaver: An autonomous evolutionary agent for exploring the programmatic space of eeg analysis pipelines. *arXiv preprint arXiv:2602.13473*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ruiqi Yang and Eric Modesitt. 2023. Vit2eeg: leveraging hybrid pretrained vision transformers for eeg data. *arXiv preprint arXiv:2308.00454*.

Sha Zhao, Mingyi Peng, Haiteng Jiang, Tao Li, Shijian Li, and Gang Pan. 2025. Eegagent: A unified framework for automated eeg analysis using large language models. *arXiv preprint arXiv:2511.09947*.

Appendix

A Experimental Details

All experiments were executed on an NVIDIA RTX 5060 GPU and an Intel Core processor. Software dependencies include Python 3.12, scikit-learn 1.4, and numpy. Classification endpoints were fully self-contained using a fixed random seed (42) to guarantee strict reproducibility across training/test splits, model initializations, and evaluation metrics. In training the baselines, standard hyperparameter selections were employed (e.g., standard ensemble counts for Gradient Boosting and Random Forest are set to 200, max depth set at 20 for RF and 5 for GB). The total compute time for agent iterations and baseline extractions took approximately several minutes per run, owing to the lightweight dataset properties.

B Complete Feature List

The LLM-guided agent selected a total subset of 40 features out of the 178 initial timeline points. The complete list of features elected by the best-performing iteration ($k = 1$) is given below.

- **Overlap with Top-39 MI Filter (39 features):** X1, X2, X3, X4, X14, X16, X31, X37, X38, X39, X44, X47, X52, X53, X67, X68, X77, X90, X91, X105, X111, X112, X124, X126, X127, X128, X129, X141, X156, X157, X158, X159, X160, X161, X162, X166, X167, X172, X174.
- **Unique to the LLM agent (1 feature):** X61.

C Per-Iteration Analysis

Over the 5 distinct interaction rounds, the agent successfully submitted distinct feature groupings, heavily influenced by temperature schedules and dynamic iteration histories.

- **Iteration 1 ($T = 0.3$):** 40 features. Reached an F1-score of 0.9107 and Accuracy of 96.48%. Validated as the top-performing configuration.
- **Iteration 2 ($T = 0.7$):** 23 features. Performance shifted down slightly to F1 of 0.8943.
- **Iteration 3 ($T = 0.7$):** 21 features. Showed further contraction prioritizing precision (Accuracy=95.52%, F1=0.8869).
- **Iteration 4 ($T = 0.7$):** 21 features. Stable subset retention (Accuracy=95.57%, F1=0.8879).
- **Iteration 5 ($T = 0.7$):** 20 features. Modest upward recovery reaching an F1 of 0.8947.

No deterministic fallback triggers were fired; each output strictly obeyed structural formatting requisites (i.e., valid JSON array), although the specific 30-feature length constraint was loosely interpreted by the model.

D LLM Prompt Templates

To ensure full reproducibility of the agentic feature selection process, we provide the exact text of the prompts supplied to the Large Language Model. The prompt architecture consists of a static system instruction, a dynamic statistical table, and an adaptive feedback mechanism that depends on the current iteration.

D.1 Initial Iteration Prompt

In the first iteration ($k = 1$), the agent is provided solely with the statistical summary of the features and is instructed to output an initial feature subset. The exact prompt template is as follows, where `{top_k}` is typically 30, and the tabular data is populated dynamically:

```
You are a feature selection agent for EEG seizure detection
(binary: seizure vs non-seizure).
Below are statistics for the top {top_k} features out of
178 total (X1-X178).

Feature    MI        Corr      RF_Imp    F_Score
--------------------------------------------------
[Dynamically generated statistical tabular data]

Select EXACTLY 30 features from X1-X178 for best seizure detection. Use the
    statistics above.
Output ONLY a JSON array of 30 feature names.
Example: ["X1","X2","X3",...,"X30"]
Your 30 features:
```

D.2 Refinement Iteration Prompt

For subsequent iterations ($k \geq 2$), the prompt is augmented with performance feedback from the preceding iteration's Random Forest evaluation. Notably, explicit mutation instructions are provided to encourage exploration and refinement based on empirical classification behavior:

```
You are a feature selection agent for EEG seizure detection
(binary: seizure vs non-seizure).
Below are statistics for the top {top_k} features out of
178 total (X1-X178).

Feature    MI        Corr      RF_Imp    F_Score
--------------------------------------------------
[Dynamically generated statistical tabular data]

Previous round: {num_prev} features, Accuracy={prev_acc:.4f}, F1={prev_f1:.4f}, AUC
    ={prev_auc:.4f}.
You previously selected: [{prev_feat_str}]. Top-5 most important features in that
    selection: {top5_str}
Bottom-5 least important features in that selection: {bot5_str}
You MUST select a DIFFERENT set of 30 features. Drop at least 5 features from your
    previous selection ideally the low-importance ones listed above and replace them
     with features NOT in your previous selection that have high mutual information
    or F-score from the table.

Select EXACTLY 30 features from X1-X178 for best seizure detection. Use the
    statistics above.
Output ONLY a JSON array of 30 feature names.
Example: ["X1","X2","X3",...,"X30"]
Your 30 features:
```

In instances where detailed importance feedback is unavailable, the mutation instruction gracefully defaults to a generic replacement directive: *"Try different features to improve F1. Add more features if too few, remove noisy ones if too many."* This structured prompt architecture reliably anchors the recursive reasoning loops described in Section 4.3.2.

Training Biomedical Retrievers From Large-Scale Citation Contexts

Xing David Wang and Duy Le Thanh and Ulf Leser
Humboldt-Universität zu Berlin
Department of Computer Science
{xing.david.wang,leser}@hu-berlin.de

Abstract

The MedCPT model has demonstrated that strong biomedical retrievers can be trained using proprietary PubMed search logs. In this work, we study whether freely available citation sentences are sufficient to train similarly effective models. We construct a large-scale training dataset of ~62 million citation sentence-abstract pairs extracted from PubMed Central. We train a lightweight BERT-based retriever-reranker model called CiteRec on this dataset and evaluate it across three benchmark settings: (a) the biomedical subset of BEIR for information retrieval, (b) SciRepEval for generalizable scientific document embeddings, and (c) CitancePlus, a new set of ~90 thousand citation sentence-abstract pairs for PubMed-scale citation recommendation.

We show that CiteRec performs competitively with MedCPT on the biomedical BEIR subset and outperforms it on SciRepEval. On CitancePlus, CiteRec achieves strong performance for citation recommendation over the full PubMed corpus, outperforming both MedCPT and a substantially larger Qwen3-Embedding-8B retriever.

1 Introduction

Dense retrievers have become increasingly effective in first-stage information retrieval (IR) as shown on several IR benchmarks (Thakur et al., 2021; Muennighoff et al., 2023; Enevoldsen et al., 2025). They also have been successfully adopted to the biomedical domain (Jin et al., 2023; Sinha et al., 2025). Dense retriever are trained using query-document relevance pairs, which can be obtained either from human-annotated datasets such as MS MARCO (Bajaj et al., 2016) or BioASQ (Tsatsaronis et al., 2015), or automatically mined pairs like search engine logs (Jin et al., 2023; Rekabsaz et al., 2021). Automatically mined relevance pairs are much easier to obtain, however, they may lack in quality compared to manually annotated pairs. Citation graphs and citation edges provide another potential source for automatically mining relevance pairs (Cohan et al., 2020; Medić and Snajder, 2020; Li et al., 2023), either using abstract-abstract, or citation sentences-abstract pairs. In the following, we call the sentence surrounding an in-text citation a "citance" (see examples in Table 1). These citances are attractive supervision signals because they often summarize or contextualize the cited work, representing potential queries answerable through the referenced document.

Among these sources, search logs have proven particularly effective in training dense retrievers. The MedCPT model (Jin et al., 2023) has demonstrated that strong biomedical retrievers can be trained using only PubMed search logs, highlighting the effectiveness of large-scale weak supervision but relying on data that is not publicly available. In this work, we study whether freely available citances alone can provide sufficient supervision to train similarly effective models for both citation recommendation and general biomedical information retrieval. Prior work leveraging citances as training data (Medić and Snajder, 2020; Gu et al., 2022) mainly focused their evaluation on the citation recommendation task. In contrast, we also evaluate their effectiveness for general biomedical retrieval. For training our model, we use the OPCitance dataset (Hsiao and Torvik, 2023) which is derived from all citances and their referenced PubMed IDs in PubMed Central full-text articles published until May 2019. After pre-processing the OPCitance dataset, our resulting dataset contains about 62 million citance-abstract pairs.

We train a lightweight BERT-based retriever-reranker model, CiteRec, on our dataset of citance-abstract pairs. Despite relying solely on freely available and partially noisy citation data, CiteRec achieves competitive performance with MedCPT on biomedical information retrieval in BEIR, outperforms it on general scientific embedding tasks

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 75–83
July 3-4, 2026 ©2026 Association for Computational Linguistics

Query	Relevant Document
The macroscopic appearance of the desmoid tumor is a pale surface with scant vascularization.	(Janssen et al., 2017)
Carboxylation is a common strategy in the assimilation of compounds that lack a functionalizable terminal carbon group.	(Erb, 2011)
High-grade tumors and pre-operative urinary cytologypositive cases are more likely to have recurrence.	(Yamada et al., 2009)

Table 1: Example citation sentences (citances) and their referenced documents from our training dataset.

and achieves strong results on citation recommendation. Our contributions are summarized as follows:

- We train the CiteRec model, comprised of a BERT-based dense retriever and cross-encoder reranker, on a large corpus of over 60 million citance-abstract pairs collected from PubMed Central.
- We systematically evaluate performance on a diverse set of biomedical and scientific benchmarks: (a) the biomedical subset of BEIR for information retrieval, (b) SciRepEval for generalizable scientific document embeddings, and (c) CitancePlus, a new citation recommendation dataset mined from PubMed Central after the OPCitance cutoff date, complementing it with over 90 thousand unseen citance-abstract pairs .
- We show that a lightweight BERT-based retriever-reranker trained solely on citance–abstract pairs generalizes well across tasks. The model performs competitively with MedCPT on biomedical BEIR and outperforms it on SciRepEval. On CitancePlus citation recommendation, it outperforms both MedCPT as well as substantially larger decoder-based embedding models such as the Qwen3-Embedding-8B retriever.

Our results suggest that citation sentences alone provide a surprisingly strong supervision signal for learning biomedical document representations.

The rest of the paper is structured as follows: We give an overview of Related Work in Section 2 and describe our training set and our model design in Section 3. In Section 4, we show our evaluation setup and how to reproduce our experiments. We report our results in Section 5 and discuss ablation studies in Section 6 before concluding our work in Section 7.

2 Related Work

Scientific citations have been widely used as training data for generalizable document embedding models. Prior work such as SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022) learn document representations from abstract–abstract citation pairs derived from the global citation graph. To keep only high-quality abstract pairs, (Liang et al., 2025) extend this approach by using the corresponding local citation sentence to filter out potentially noisy pairs before large language model training. In contrast to these approaches, our model directly trains on citance-abstract pairs instead of indirect abstract-abstract supervision.

Leveraging citances for citation recommendation has been introduced by He et al. (2010). Current models typically follow an IR workflow, consisting of a first-stage retriever before reranking the retriever results (Gu et al., 2022). Newer approaches (Medić and Snajder, 2020; Gu et al., 2022) have incorporated the article title and abstract to the query in addition to the citance for a richer flow of information. In our work, we want to examine how well such models perform not only on citation recommendation but also on general IR queries such as found in BEIR (Thakur et al., 2021). Thus, we stick to shorter, asymmetrical queries consisting of citances only.

Dense IR models are usually trained on large-scale annotated datasets such as MS Marco (Bajaj et al., 2016) and evaluated on annotated data as well, such as the BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) benchmarks. In the biomedical domain, however, annotated data is scarce motivating alternative sources. A key contri-

bution in this direction is MedCPT (Jin et al., 2023) which demonstrated that training biomedical retrievers solely on large-scale weak supervision data from PubMed search and click logs achieves strong performance on biomedical IR benchmarks. MedCPT establishes a strong baseline for biomedical retrieval using weak supervision, but relies on proprietary data. In contrast, we use citance-abstract pairs mined from PubMed Central full-text articles as a fully open alternative source of supervision, enabling large-scale training without privacy constraints.

Besides domain-specific fine-tuning, recent approaches (Li et al., 2023; Xiao et al., 2023; Zhang et al., 2025) have also explored scaling dense retrievers through larger models and larger training datasets. The gte model family (Li et al., 2023) employs a pre-training step by training first on automatically mined query-document pairs before fine-tuning on high-quality, manually annotated retrieval pairs. Among the pre-training data, they also employ a set of citance-abstract pairs from the S2ORC corpus (Lo et al., 2020) as auxiliary supervision signal. (Sinha et al., 2025) continue fine-tuning the gte models on pairs of document abstracts and synthetic queries targeting those. To assess how weak supervision from one source compares to these larger models, we also compare our approach against these embedding models.

In contrast to prior work, we investigate whether large-scale citance–abstract pairs alone are sufficient to train biomedical retrievers that generalize across both citation recommendation and general information retrieval tasks.

3 Methods

3.1 CiteRec Overview

We first explain our methodology to obtain biomedical citances before describing the two main components of our CiteRec model: (i) the lightweight dense retriever to quickly obtain an initial set of candidate documents and (ii) the cross-encoder reranker to refine the ranking of the candidates.

3.2 Training Data

For training, we use the OPCitance dataset created by Hsiao and Torvik (2023) who extracted all pairs of citances and referenced documents from PubMed Central until the cutoff date of May 2019. Compared to other full-text corpora such as S2ORC (Lo et al., 2020), OPCitance is solely focused on the biomedical domain. In our setup, citances function as queries while the title and abstract of references serve as the documents to be retrieved. We employ two steps for data pre-processing: First, we filter out all documents that were cited but were not assigned any valid PubMed ID as the reference. Secondly, we filter out citation contexts longer than 64 tokens discarding long queries (i) with potentially malformed or irrelevant context and (ii) to ensure that the reranker obtains sufficient document context (we allow up to 512 tokens for the combination of citation and reference document). After preprocessing, we retain roughly 62M citance-abstract pairs and show some example citance-abstract pairs in Table 1.

3.3 Retriever

For our CiteRec retriever, we train separate query and document encoders both initialized from a PubMedBERT-base model (Gu et al., 2021). We take the final hidden state of the [CLS] token as the respective query and document embeddings. For retriever training, we optimize a query-to-document loss function

$$\ell_i^{q2d} = -\log \frac{\sum_{k \in Pos(q_i)} sim(q_i, d_k)}{\sum_{m=1}^{|B|} sim(q_i, d_m)},$$

whereas $|B|$ denotes the batch size. q_i is a citance query, d_i its associated positive document and the similarity $sim(q, d)$ is implemented by the dot product of both embeddings. Extending the standard InfoNCE loss formulation (Oord et al., 2018), we allow multiple positive documents $d_k \in Pos(q_i)$ for a given query q_i. For negatives, we use other citation pairs in the same batch as in-batch negatives. However, instead of relying on batching random pairs together, we group citation pairs from the same full text into the same batch. This allows us to mine potentially more informative negatives as documents referenced by the same paper are more likely to be thematically related, hence acting as harder negatives than other, random documents.

Similarly to MedCPT, we also define a complementary document-to-query loss ℓ_i^{d2q} which models relevance of individual queries to a given document d_i. We combine both losses to a batch loss ℓ_B:

$$\ell_B = \frac{\alpha}{|B|} \sum_{1}^{|B|} w_i \ell_i^{q2d} + \frac{1-\alpha}{|B|} \sum_{1}^{|B|} w_i \ell_i^{d2q}$$

where the hyperparameter $\alpha = 0.8$ controls the influence of each individual loss term and the weight w_i models the importance of citing paper and cited references according to their citation counts. To normalize citation counts across fields, we use the Relative Citation Ratio rcr[1] (Hutchins et al., 2016) and set the weights $w_i = \log(1+rcr_i)$. The rcr value takes into account the year of a publication and its citation numbers relative to its field.

3.4 Reranker

For CiteRec reranker training, we initialize a cross-encoder from the PubMedBERT-base model (Gu et al., 2021). The cross-encoder optimizes the same InfoNCE loss ℓ_i^{q2d} as the retriever but uses random in-batch negatives as the reranker can leverage random negatives more effectively due to its cross-attention mechanism. It takes as input the sequence [CLS] q_i [SEP] d_i [SEP] with the query q_i, the document d_i and the special BERT tokens [CLS] and [SEP], and calculates the similarity score

$$sim(q, d) = W^T e([CLS]) + b$$

where e([CLS]) is the final hidden state embedding of the [CLS] token and $W^T \in \mathbb{R}^d, b \in \mathbb{R}$ are learnable weights. The reranker takes into account the top-100 documents returned by our retriever.

4 Experiments

4.1 Evaluation Datasets

We conduct our experiments on three diverse biomedical IR and document embedding benchmarks.

BEIR (Thakur et al., 2021) is a zero-shot information retrieval benchmark to assess the generalization capabilities of newly developed retrieval models. Similarly to (Jin et al., 2023), we evaluate our model on the same subset of five BEIR tasks as (Jin et al., 2023) which include TREC-COVID (Voorhees et al., 2021), containing questions related to the COVID-19 pandemic, NFCorpus (Boteva et al., 2016), questions regarding nutrition facts, BioASQ (Tsatsaronis et al., 2015), a general-purpose biomedical question answering dataset over the entirety of PubMed, SciFact (Wadden et al., 2020), retrieving supporting documents that verify scientific claims and SciDocs (Cohan et al., 2020), finding similar documents given the title and abstract of a paper.

SciRepEval (Singh et al., 2023) is aimed at assessing the quality of document embedding models and includes tasks besides retrieval/search, such as classification and regression. Exemplary classification tasks include assigning documents to a MeSH descriptor and to its field of study. Exemplary regression tasks include predicting the citation count and the year of a publication. SciRepEval distinguishes between tasks whose datasets were used during the training of their embedding models (in-train) and tasks that were not (out-of-train, none of our evaluated models were trained on any of the in-train datasets.).

Our third benchmark deals with the task of recommending references for given citation contexts from PubMed. For this, we create a new, independent test set called CitancePlus mined from PubMed Central full-text articles published after May 2019 to prevent overlaps with the OPCitance dataset used for our model training. To obtain query-document pairs, we randomly sampled 10,000 full texts published between June 2019 and June 2024 and for each of them, sampled up to 10 in-line citations as queries. We then extract all referenced PubMed articles of each citance as valid reference documents. The resulting dataset contains 32,000 query-document pairs in the development split and 93,000 pairs in the test split. We used the official PubMed baseline of 2025 as the document corpus (containing all PubMed abstracts until January 2025).

4.2 Baselines

We compare the results of our CiteRec model to several baselines: BM25 (Robertson and Zaragoza, 2009) is a lexical retriever serving as a strong baseline in IR tasks (Thakur et al., 2021). MedCPT (Jin et al., 2023) is the baseline closest to ours, training a dual bi-encoder setup with a separate query and document embedder initialized from PubMedBERT-base, while also training a cross-encoder reranker on top. While we exclusively train on freely available citance-abstract pairs, MedCPT exclusively trains on a large proprietary set of 255 M PubMed search logs containing pairs of user query and abstract[2].

[1]We download the most recent dump from March 2025 under https://nih.figshare.com/articles/dataset/iCite_Database_Snapshot_2025-03/28789178.

[2]We use the MedCPT models provided via Huggingface (https://huggingface.co/ncbi/MedCPT-Article-Encoder) which differ a bit from the results reported in the original paper.

SciNCL (Ostendorff et al., 2022) and SPECTER (Cohan et al., 2020) are document embedding models that both leverage the citation graph for training similar to our approach. Note that they are using abstract-abstract pairs instead of citances for training and thus have not been explicitly designed for short search queries.

The last group of baselines are retriever models trained on a much larger mixture of data including search logs, web links and manually annotated pairs of questions and answer documents: (i) BERT-based encoder models such as the gte family (Li et al., 2023) and bge (Xiao et al., 2023) have around 100 to 300 million model parameters resulting in a similar size as MedCPT and CiteRec but training on more diverse retrieval pairs; the gte models use around 800 million retrieval pairs during unsupervised pre-training and around three million high-quality pairs for additional fine-tuning, the bge models use a similar pipeline of pre-training and subsequent fine-tuning with around 200 million pairs. (ii) Decoder-only embedding models like Qwen3 (Zhang et al., 2025) are much larger in size with around 8 billion model parameters and trained on a carefully selected set of around 150 million retrieval pairs during pre-training and 12 million pairs during fine-tuning.

4.3 Implementation Details

We conducted training for the retriever on two NVIDIA A100 GPUs with a batch size of 128, gradient accumulation step size of 4, using bf16 precision, a cosine scheduler with a linear warm-up phase, and a target learning rate of 2e-6. We trained the model for one epoch resulting in a runtime of around four days. As cross-encoder reranker training converges much faster, we applied early stopping on the development set to find our best model there. We set the batch size for the reranker training to 1, the gradient accumulation step size to 32 and the target learning rate of 5e-6. The CitancePlus corpus, the CiteRec retriever, the CiteRec reranker and the accompanying code is available at under `https://github.com/WangXII/CiteRec`.

5 Results

In Table 2, we report the results of our CiteRec model and of the baselines on the subset of biomedical tasks in BEIR. As expected, even though the SciNCL model was trained on abstract-abstract citation links, it does not perform well on the biomedical BEIR subset except for SciDocs. Our CiteRec retriever is better suited for IR queries than SciNCL by learning from citance-abstract pairs instead.

Looking at individual subtasks and retriever-only models, our CiteRec retriever achieves particularly good results on the SciFact dataset with an NDCG@10 score of 0.765 outperforming all other first-stage retrievers. On average across all subtasks, our CiteRec retriever performs competitively against the MedCPT retriever (0.425 NDCG@10 compared to 0.431) showing that training on public citation data alone can match up to training on proprietary search logs. Adding a reranker on top of the CiteRec retriever improves retrieval performance on all examined datasets resulting in an average increase of around 5.3 percentage points (pp). However, here, bigger gaps are observed on the BioASQ and TREC-COVID datasets when compared to the corresponding MedCPT reranker indicating that real-world user queries might help more in reranker training for biomedical question answering tasks. Compared to the gte and bge models which trained on a richer mixture of (high-quality) data, the MedCPT and CiteRec retrievers trained only on one (noisy) source of data show lower performances. However, by adding the rerankers, both models close the gap with CiteRec achieving comparable performances and MedCPT even surpassing them.

When evaluating quality of document embeddings on the SciRepEval benchmark (see Table 3), we can see that our CiteRec retriever is able to outperform the MedCPT retriever by over 3.3 pp (also reflecting the more citation-focused subtasks of SciRepEval). As expected, the SPECTER and SciNCL models perform well on this task as they are designed to jointly optimize document abstract embeddings. SciNCL achieves an average score of 0.688 and SPECTER a score of 0.675 compared to 0.673 for CiteRec. Still, the bge-large model achieves the best results overall with a score of 0.694. CiteRec performs competitively overall with the models trained on a larger set of data, achieving the joint second best results on the In-Train part of SciRepEval together with gte-large (both 0.559).

Our last main experiment focuses on the CitancePlus dataset ranking relevant abstracts in PubMed given a local citation context. We present the results on CitancePlus in Table 4 measured by NDCG and MAP scores @10 and @100. We observe that BM25 is a strong baseline for this task outperforming neural models like gte-base and Med-

	Param Size	TREC-COVID	NFCorpus	BioASQ	SciFact	SciDocs	Avg
BM25	-	0.656	0.325	0.465	0.665	0.158	0.454
gte-base	100 M	0.676	0.374	0.355	0.753	0.229	0.477
gte-large	300 M	0.687	**0.381**	0.358	0.750	**0.237**	0.482
bge-large	300 M	**0.744**	0.343	0.400	0.736	0.218	0.488
SciNCL	100 M	0.342	0.228	0.108	0.568	0.194	0.288
MedCPT	100 M	0.616	0.365	0.325	0.736	0.114	0.431
+ reranker	200 M	0.727	0.365	**0.498**	**0.793**	0.176	**0.512**
CiteRec (Ours)	100 M	0.601	0.299	0.314	0.765	0.148	0.425
+ reranker (Ours)	200 M	0.675	0.347	0.392	0.792	0.184	0.478

Table 2: Biomedical BEIR results measured by NDCG@10. Best results in each column in bold, second best underlined.

	In-Train	Out-of-Train	Avg
gte-large	0.559	0.720	0.683
bge-large	**0.575**	**0.736**	**0.694**
SPECTER	0.547	0.720	0.675
SciNCL	0.556	0.734	0.688
MedCPT	0.531	0.678	0.640
CiteRec (Ours)	0.559	0.713	0.673

Table 3: SciRepEval Results. Best results in each column in bold, second best underlined.

CPT across all metrics and gte-large in all but the NDCG@100 score (0.193 compared to 0.201). The MedCPT retriever struggles with the citation recommendation task only reaching an NDCG@10 of 0.097. Our CiteRec retriever model outperforms all other similarly sized retriever models on this task by more than 3 pp across all metrics, showing that the citation recommendation task benefits a lot from domain-specific finetuning.

We compare to the Qwen3 model only on the citation recommendation task as the gte and bge models have already proven to be challenging baselines on the biomedical BEIR and SciRepEval benchmarks and evaluating large decoder-based embedding models at PubMed scale is computationally expensive. On CitancePlus, the Qwen3-Embedding-8B model slightly outperforms our CiteRec first-stage retriever with an NDCG@10 score of 0.216 compared to 0.206. However, after adding the CiteRec reranker on top of the retriever, our overall CiteRec model even outperforms the Qwen3-Embedding-8B model by 2.4 pp in NDCG@10 score while being significantly more efficient in terms of parameter size and inference cost.

6 Discussion

In our first ablation study, we examine how retrieval performance of our CiteRec retriever changes when we modify our negative sampling from grouped batching of negatives from the same publication to random in-batch negatives and replacing the rcr values as individual weights for a document by the default value of 1. We report the corresponding results in Table 5. We observe that grouped batching improves NDCG@10 scores on both datasets, with the improvements being more pronounced on the CitancePlus dataset (0.206 vs 0.164) compared to the biomedical subset of BEIR (0.425 vs 0.412). This suggests that retriever training likely benefits from harder examples coming from the same documents whereas random negatives are easier and saturate model performance more quickly.

Our second ablation study examines the capabilities of our CiteRec models in a stricter setting: We constructed our original CitancePlus test dataset in such a way that no test query overlaps with any query encountered in the training set. However, the model might still have incorporated a slight bias toward documents already encountered in the OPCitance training data. To minimize such potential effects, we repeated the CitancePlus experiments in a stricter setting removing all citance-abstract pairs from the test dataset where the abstract document has already occurred at least once in the training dataset. We reduced the initial 93 thousand citance-abstract pairs to a subset of around 25 thousand after applying this filtering step.

Table 6 summarizes the results on the filtered test dataset. We see that the performance of BM25 increases on the filtered test dataset from 0.161 NDCG@10 to 0.185 NDCG@10. Our CiteRec

	Param Size	NDCG@10	NDCG@100	MAP@10	MAP@100
BM25	-	0.161	0.193	0.132	0.139
gte-base	100 M	0.143	0.178	0.121	0.127
gte-large	300 M	0.157	0.201	0.125	0.134
bge-large	300 M	0.170	0.212	0.137	0.145
Qwen3-Embedding-8B	8 B	<u>0.216</u>	<u>0.268</u>	<u>0.175</u>	<u>0.186</u>
SciNCL	100 M	0.076	0.106	0.057	0.063
MedCPT	100 M	0.097	0.132	0.075	0.082
+ reranker	200 M	0.174	0.191	0.146	0.149
CiteRec (Ours)	100 M	0.206	0.249	0.168	0.177
+ reranker (Ours)	200 M	**0.240**	**0.276**	**0.200**	**0.209**

Table 4: CitancePlus results. Best results in each column in bold, second best underlined.

	CitancePlus	BEIR
CiteRec	0.206	0.425
CiteRec (BatchNeg)	0.164	0.412

Table 5: Ablation on our CiteRec retriever model to train only using random in-batch negatives. Resulted scores are NDCG@10.

CitancePlus	Test	Filtered
BM25	0.161	0.185
gte-large	0.157	0.162
Qwen3-Embedding-8B	<u>0.216</u>	<u>0.221</u>
CiteRec (Ours)	0.206	0.205
+ reranker (Ours)	**0.240**	**0.248**

Table 6: Ablation on the CitancePlus dataset removing any paper from the test dataset which was also cited at least once in the training data. Result scores in NDCG@10. Best results in each column in bold, second best underlined.

retriever decreases only minimally in NDCG@10, from 0.206 to 0.205, indicating at most a slight bias toward documents already seen in the OPCitance training data. The larger retriever models gte-large and Qwen3-Embedding-8B improve by 0.5 pp in NDCG@10 score. However, when adding our reranker model to the top-100 retrieved documents, overall NDCG@10 score improves again by 0.8 pp compared to the original test dataset indicating that any such bias is not as pronounced in the reranker model.

7 Conclusion

In this work, we introduced a biomedical retriever-reranker model trained on a large set of PubMed Central citation-abstract pairs. We have shown that the resulting model generalizes well on general biomedical IR benchmark such as the BEIR subset when compared to models of similar size and training data such as MedCPT and outperforms both these baselines as well as substantially larger state-of-the-art models (e.g., Qwen3-Embedding-8B) when compared on the in-domain citation recommendation task.

Limitations

Training on citance-abstract pairs is inherently noisy as citations do not always match what is written in the referenced texts and not all information referred to in a given context is fully included in the title and abstract of the referenced document (Sarol et al., 2024). We did not further investigate any methods to filter out high-quality citations from lower-quality ones.

We also limited our experiments to smaller BERT-based model (100M parameters) and did not

train on larger decoder-only retrievers due to limited computational resources. It remains unclear whether scaling to larger models would improve performance or lead to overfitting on the noisy training data.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft [RTG2424].

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2270–2282.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Tobias J Erb. 2011. Carboxylases in natural and synthetic microbial pathways. *Applied and environmental microbiology*, 77(24):8466–8477.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and SciBERT-based reranking. *Preprint*, arxiv:2112.01206 [cs].

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430.

Tzu-Kun Hsiao and Vetle I. Torvik. 2023. OpCitance: Citation contexts identified from the PubMed central open access articles. 10(1):243.

B. Ian Hutchins, Xin Yuan, James M. Anderson, and George M. Santangelo. 2016. Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. 14(9):e1002541.

ML Janssen, DLM Van Broekhoven, JMM Cates, WM Bramer, JJ Nuyttens, Alessandro Gronchi, S Salas, S Bonvalot, DJ Grünhagen, and C Verhoef. 2017. Meta-analysis of the influence of surgical margin and adjuvant radiotherapy on local recurrence after resection of sporadic desmoid-type fibromatosis. *Journal of British Surgery*, 104(4):347–357.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arxiv:2308.03281 [cs].

Yuan Liang, Massimo Poesio, and Roonak Rezvani. 2025. Beyond citations: Integrating finding-based relations for improved biomedical article representations. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 297–306. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Zoran Medić and Jan Snajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 97–103. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688. Association for Computational Linguistics.

Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. Tripclick: the log files of a large health web search engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2513.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

M Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. 2024. Assessing citation integrity in biomedical publications: Corpus annotation and NLP models. page btae420.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566. Association for Computational Linguistics.

Aarush Sinha, Pavan Kumar S, Roshan Balaji, and Nirav Pravinbhai Bhatt. 2025. BiCA: Effective biomedical dense retrieval with citation-aware hard negatives. *Preprint*, arxiv:2511.08029 [cs].

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yoshiaki Yamada, Yasusuke Inoue, Kogenta Nakamura, Katsuya Naruse, Shigeyuki Aoki, Tomohiro Taki, Motoi Tobiume, Kenji Zennami, Remi Katsuda, Kouji Hara, and 1 others. 2009. Long-term results and management of ureteral transitional cell carcinoma using the holmium: Yag laser via rigid-ureteroscopy. *Oncology reports*, 21(2):345–349.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arxiv:2506.05176 [cs].

Reliable Automated Triage in Spanish Clinical Notes: A Hybrid Framework for Risk-Aware HIV Suspicion Identification

Rodrigo Morales-Sánchez[1,*], Soto Montalvo[2], Raquel Martínez[1],

[1]Dept. Lenguajes y Sistemas Informáticos, ETSI Informática,
Universidad Nacional de Educación a Distancia (UNED)
[2]Dept. Informática y Estadística, ETSI Informática,
Universidad Rey Juan Carlos (URJC)
[1]{*rmorales,raquel}@lsi.uned.es, [2]soto.montalvo@urjc.es

Abstract

Standard clinical Natural Language Processing (NLP) benchmarks often yield inflated metrics by forcing deterministic classification on ambiguous instances, thereby obscuring the clinical risks of overconfident predictions. To bridge this gap, we propose a risk-aware hybrid selective classification framework, evaluated on early Human Immunodeficiency Virus suspicion identification in Spanish clinical notes. Our dual-verification approach explicitly decouples aleatoric uncertainty through Mondrian conformal prediction and epistemic uncertainty using a Multi-Centroid Mahalanobis Distance veto. Empirical evaluations reveal that standard uncertainty metrics and baseline classifiers are structurally insufficient for safe medical triage, suffering severe coverage collapse when forced to operate under strict reliability constraints. In contrast, by demanding that clinical narratives pass both probabilistic and geometric safeguards, the proposed framework successfully isolates a highly trustworthy operational domain. The obtained results show that explicit, decoupled uncertainty quantification is essential for translating biomedical NLP into responsible clinical practice.

1 Introduction

The early diagnosis of Human Immunodeficiency Virus (HIV) remains a critical public health challenge. While automated Natural Language Processing (NLP) systems offer a scalable mechanism to triage Electronic Health Records (EHRs) and allocate diagnostic resources by filtering out clear negative cases (Feller et al., 2018; Morales-Sánchez et al., 2024), their transition into clinical practice is hindered by rigid evaluation paradigms.

The prediction task identifies documented clinical suspicion of infection. The supervision signal reflects clinical suspicion evidenced in text rather than confirmed serological status. Since documentation varies in completeness and specificity, suspicion is a subjective, partially observable construct. Consequently, linguistic ambiguity and distributional heterogeneity are intrinsic properties of this problem. When standard deterministic classifiers are forced to make rigid binary decisions on such narratives, they routinely produce overconfident misclassifications, directly conflicting with the safety requirements of clinical medicine (Kelly et al., 2019; Guo et al., 2017).

To translate biomedical NLP research into responsible practice, AI systems must shift from forced classification to action-oriented selective screening via the classification with rejection paradigm. However, standard Uncertainty Quantification (UQ) methodologies, such as Bayesian approximations or deep ensembles, are structurally insufficient for highly imbalanced clinical text (Abdar et al., 2021). By relying solely on predictive entropy, they conflate distinct sources of risk, causing models to either indiscriminately defer true positive cases or unsafely automate complex, out-of-distribution (OOD) narratives.

To address standard uncertainty estimation vulnerabilities, we propose a post-hoc Hybrid Selective Screening Framework that decouples aleatoric and epistemic uncertainties. Our system enforces a dual-verification policy: it utilizes Mondrian Conformal Prediction (MCP) to manage probabilistic ambiguity and a Multi-Centroid Mahalanobis Distance (MCMD) veto within a spectrally normalized latent space to filter epistemic anomalies. This intersectional safeguard translates risky forced-binary predictions into safe, trinary triage recommendations.

The main contributions of this work are:

- **Hybrid Selective Screening Framework:** We develop a novel, dual-veto deferral system that explicitly combines conformal prediction for aleatoric ambiguity with a geometrically constrained feature space for epistemic

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 84–96
July 3-4, 2026 ©2026 Association for Computational Linguistics

anomaly detection. We show that the strict intersection of both safeguards is essential for safe clinical automation.

- **Subjective Construct Modeling in Spanish EHRs:** We formalize the extraction of early HIV clinical suspicion as a partially observable inference task, addressing the linguistic and epistemological challenges of clinical NLP in a resource-constrained, non-English environment.

- **Rigorous Clinical Validation:** Utilizing asymmetric, risk-aware clinical metrics (e.g., Custom Risk-Kappa, True Positive Deferral Rate), we empirically expose the operational failure of standard UQ algorithms on clinical data. We prove that our proposed hybrid policy significantly outperforms standard baseline UQ methods, maintaining robust calibration without sacrificing triage efficiency.

2 Related Work

Early HIV Prediction via NLP Historically, automated HIV risk assessment relied on structured EHR fields, which consistently underperform as stigmatized behavioral risks are rarely captured in standard coding (Feller et al., 2018). Consequently, research has pivoted toward NLP to extract nuanced indicators from unstructured narratives using contextual word embeddings (Morales-Sánchez et al., 2024; Sah et al., 2025). As these predictive architectures mature, there is a consensus that deterministic models require explicit reliability frameworks for safe clinical adoption (Ngema et al., 2026).

Reliability and Selective Classification The deployment of AI in high-stakes public health settings requires rigorous UQ (Abdar et al., 2021). Recent work has applied selective classification to clinical text, demonstrating that deep learning classifiers can achieve strict safety targets by abstaining from unreliable samples without the prohibitive computational burden of model retraining (Peluso et al., 2024; Garcia et al., 2026). However, standard UQ methodologies struggle with highly imbalanced clinical text because their reliance on predictive entropy conflates aleatoric and epistemic uncertainty (Ziletti and D'Ambrosi, 2026). In medical triage, this conflation is operationally fatal, leading to indiscriminate deferrals or unsafe automation (Toure and Stephens, 2026). Because safe clinical deployment demands models that explicitly know when to abstain (Machcha et al., 2025), there is a pressing need for hybrid architectures that decouple and independently bound these distinct predictive risks.

3 Material and Methods

3.1 Hybrid Selective Screening and Deferral Policy

We implement a post-hoc Dual-Verification Selective Classification that requires a text to pass two distinct uncertainty checks before permitting an automated triage decision.

Aleatoric Control via Mondrian Conformal Prediction (MCP) To manage aleatoric uncertainty arising from inherent ambiguity in the clinical narratives, we employ MCP (Boström et al., 2021). First, we apply Temperature Scaling (Guo et al., 2017) to the validation logits to minimize the Negative Log Likelihood (NLL) and mitigate overconfidence. Subsequently, MCP generates a prediction set $\Gamma^{\alpha}(x) \subseteq \{0, 1\}$ containing the true label with probability $1 - \alpha$. We use the set's cardinality to indicate ambiguity: sets containing multiple classes ($|\Gamma^{\alpha}(x)| > 1$) or zero classes ($|\Gamma^{\alpha}(x)| = 0$) signify insufficient or contradictory evidence at the target risk tolerance α, necessitating a deferral.

Epistemic Control via Geometric Veto Deep neural networks frequently suffer from the phenomenon of assigning high softmax probabilities to anomalous, OOD inputs. To mitigate this, we integrate a geometric veto mechanism that operates directly within the L_2-normalized latent feature space. Because clinical "suspicion" encompasses highly heterogeneous patient presentations—ranging from varying risk behaviors to sparse opportunistic infections—assuming a unimodal distribution for the feature space leads to artificially inflated variances. To accurately accommodate this phenotypic diversity without risking the severe overparameterization of local density estimation in a high-dimensional space, we model the clinical narratives using a MCMD framework (effectively a tied-covariance Gaussian Mixture).

For each class c, the optimal number of local centroids, K_c, is dynamically determined via k-means clustering by identifying the smallest K where the relative inertia gain falls below a predetermined 0.05 threshold, subject to a minimum cluster size constraint. Rather than estimating independent, unstable covariance matrices for every local clus-

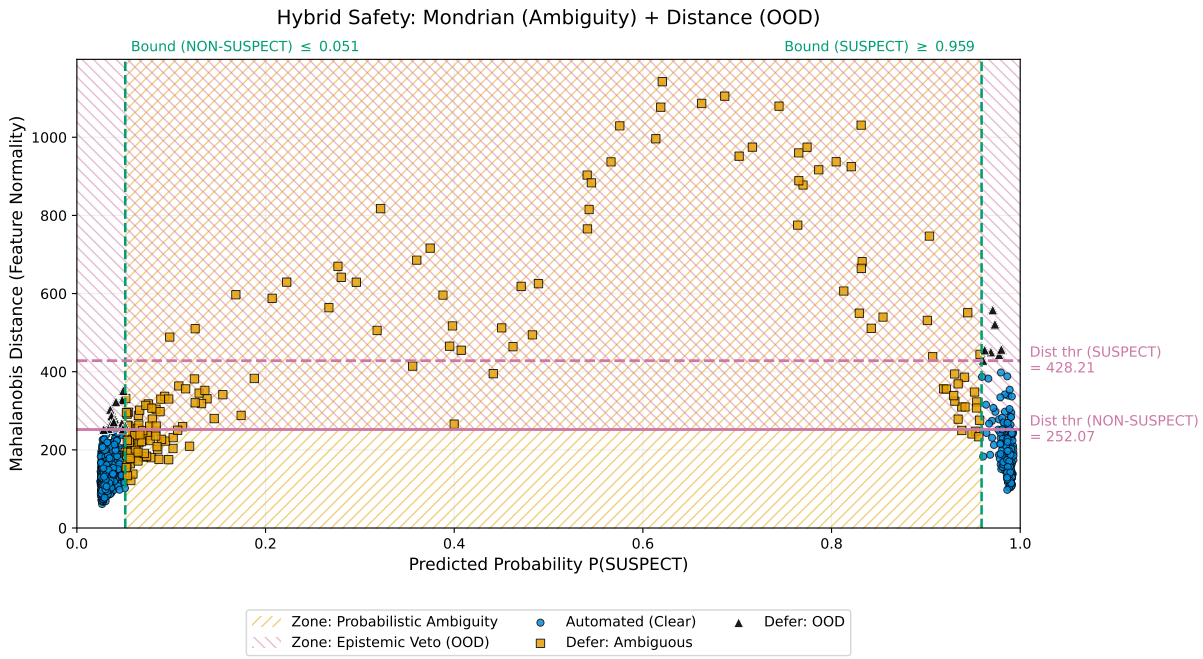


Figure 1: Hybrid Selective Screening Policy in the spectrally normalized latent space. Soft teal vertical dashed lines denote MCP boundaries for probabilistic ambiguity, mapped to the diagonally hatched zone. Reddish-purple horizontal lines indicate Mahalanobis distance thresholds (τ_{dist}) for the epistemic veto, mapped to the cross-hatched zone. Automated 'Clear' cases are shown as medium blue circles, probabilistically ambiguous deferrals as golden orange squares, and epistemic OOD deferrals as solid black triangles.

ter, we assume a globally shared noise structure. By pooling the residuals of all samples to their assigned centroids, a single, robust global precision matrix Σ^{-1} is fitted across all classes using Oracle Approximating Shrinkage (OAS).

During inference, the epistemic uncertainty of a sample x is quantified as the minimum Mahalanobis distance between its embedding $f(x)$ and any centroid $\mu_{\hat{y},k}$ associated with its predicted class $\hat{y}$:

$$d_M(x, \hat{y}) = \min_{k \in \{1 \dots K_{\hat{y}}\}} \sqrt{r_{\hat{y},k}^\top \Sigma^{-1} r_{\hat{y},k}}$$

where $r_{\hat{y},k} = f(x) - \mu_{\hat{y},k}$ represents the residual vector between the sample embedding and the k-th centroid.

To account for the asymmetric variance between routine (Class 0) and highly heterogeneous suspected cases (Class 1), we establish class-specific distance thresholds $\tau_{dist}^{(\hat{y})}$, calibrated to the 99th percentile of intra-class distances from correctly classified validation samples. Inputs exceeding this boundary are flagged as geometrically anomalous. This mechanism effectively vetoes overconfident OOD predictions without disproportionately deferring the diverse minority class.

These specific thresholds were not chosen arbitrarily; a post-hoc grid search over the validation folds confirmed that the 0.05 inertia and 99th percentile boundaries empirically maximized the Custom Risk-Kappa, optimally balancing OOD anomaly detection with minority class variance (see Appendix A for the full sensitivity analysis).

Global Decision Logic The final deferral policy $D(x)$ enforces a strict intersection of both safeguards:

$$D(x) = \begin{cases} \hat{y} & \text{if } |\Gamma^\alpha(x)| = 1 \\ & \text{AND } d_M(x, \hat{y}) \leq \tau_{dist}^{(\hat{y})} \\ \perp & \text{otherwise} \end{cases}$$

A complete visual representation of this dual-verification logic is provided in Figure 1.

3.2 Data

The study utilizes a proprietary cohort of Spanish EHRs sourced from Hospital Universitario Fundación Alcorcón (HUFA, Madrid), comprising real-world clinical notes. The total collection comprises 13,642 patients with notes between 1998 and 2023 from different medical specialities. All data was anonymized at the source by the hospital in strict compliance with the GDPR and Spanish

Metric	Total	Train (Avg.)	Validation (Avg.)	Test (Held-out)
Total Patients	13,642	11,046	1,223	1,373
Total Suspects	**2,072**	**1,671**	**183**	**218**
HIV+ / HIV-	*172 / 1,900*	*136 / 1,535*	*14 / 169*	*22 / 196*
Non-Suspects	11,570	9,375	1,040	1,155
Class Ratio (Suspect : Non-Suspect)	1:5.5	1:5.6	1:5.7	1:5.3
Total Clinical Notes	63,802	51,642	5,710	6,450
Avg. Notes / Patient	-	4.68	4.67	4.70
Tokens / Patient (Mean / Max)	-	409 / 8,621	407 / 8,033	427 / 5,255
Avg. Tokens / Note (Susp. / Non-Susp.)	-	∼126 / ∼79	∼118 / ∼79	∼127 / ∼81

Table 1: Dataset total and partitioning with key characteristics. Training and validation figures represent the average across the 10-fold cross-validation splits.

LOPDGDD regulations by replacing 29 different entity types (including age and gender) with placeholder tags.

Rather than relying on manual annotation, ground truth was established empirically based on documented physician behavior: the suspect cohort comprises patients who had a requested HIV serology (both positive and negative outcomes). EuroTEST (EuroTEST, 2016) indicator conditions were then applied as a secondary filter to isolate genuine clinical suspicion, excluding cases where serology was requested as a routine administrative procedure. Because the threshold for ordering a test varies among physicians, this supervisory signal inherently captures the subjective, partially observable nature of real-world clinical suspicion, further motivating our selective screening paradigm. However, because clinical documentation practices and physician thresholds for ordering serology vary substantially across institutions, it is important to note that the geometric thresholds calibrated on this single-center cohort would likely require recalibration prior to external deployment.

To prevent explicit diagnostic leakage and force the model to learn the implicit digital phenotype of early HIV risk, the corpus was aggressively sanitized by removing all explicit HIV acronyms and direct testing recommendations. The non-suspect cohort included patients with no history of such requests. The data was partitioned using 10-fold cross-validation alongside a held-out test set (Table 1).

3.3 Model Architecture

3.3.1 Attention-MIL Architecture

To handle EHR structural variability (Table 1) without discarding longitudinal context, we frame HIV suspicion detection as a Multiple Instance Learning (MIL) problem, adapting Ilse et al. (2018).

In this approach, each patient's record is treated as a 'bag' of overlapping text sequences. Documents are partitioned using a sliding window strategy (up to 64 chunks of 384 tokens) and independently encoded using the Spanish biomedical RoBERTa model `PlanTL-GOB-ES/bsc-bio-ehr-es`[1] (Carrino et al., 2022). To aggregate these chunks into a unified document representation $\mathbf{z}_{doc}$, we employ a gated attention pooling mechanism. The network assigns a learned scalar weight to each chunk via a gated combination of hyperbolic tangent and sigmoid activations, ensuring that highly suspicious text snippets dominate the final document-level embedding before it is passed to the UQ evaluation.

3.3.2 Training for Reliable Uncertainty

To address the class imbalance while maintaining calibration, we contrast two architectures:

Standard MIL: Optimized using a Label-Smoothed Focal Loss (Lin et al., 2017) to down-weight easy negatives, coupled with Regularized Dropout (R-Drop) (Liang et al., 2021) to enforce KL-divergence consistency between forward passes, thereby smoothing the decision boundary.

MD-SN MIL: While Focal Loss improves raw classification, it can distort the underlying probabilities required for aleatoric UQ. Therefore, we utilize the MD-SN (Mahalanobis Distance + Spectral Normalization) architecture proposed by (Vazhentsev et al., 2022), coupled with Logit Adjustment (Menon et al., 2021). This technique adds logarithmic class priors directly to the raw logits, inherently pushing the decision boundary away from the minority class without sacrificing probabilistic integrity. To enable the epistemic geometric

[1] PlanTL-GOB-ES/bsc-bio-ehr-es - Hugginface. Accessed: [2026-02-26]

veto, SN is applied to the dense layers of the MIL head, bounding the Lipschitz constant to ensure a bi-Lipschitz smooth feature space.

3.4 Uncertainty Estimation Backends

To demonstrate that our selective screening framework is adaptable and to evaluate the trade-off between computational cost and empirical safety, we extract predictive uncertainty estimates using three established methodologies:

Standard UQ Backends We use Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) as a baseline Bayesian approximation, and Cross-Validation (CV) Deep Ensembles (Lakshminarayanan et al., 2017) as our empirical gold standard. The first executes T stochastic forward passes during inference to compute predictive entropy as an epistemic uncertainty threshold. The latter aggregates predictions from K models trained on distinct CV folds. Crucially, we utilize an Out-Of-Fold (OOF) calibration strategy to fit deferral thresholds strictly on unseen validation folds, preventing optimistic bias.

Mahalanobis Distance with Spectral Normalization (MD-SN) To circumvent the computational overhead of (T times) sampling or (O(K) latency and storage costs of 10 models) ensembling, we evaluate deterministic, single-pass UQ using the MD-SN architecture (Vazhentsev et al., 2022). Unlike entropy-based methods that rely on softmax dispersion, MD-SN measures uncertainty geometrically. By calculating the Mahalanobis distance between a test embedding and the fitted class centroids in a bi-Lipschitz smooth feature space, texts that are semantically atypical or OOD yield proportionally large geometric distances.

3.5 Evaluation Framework and Clinical Metrics

Standard NLP evaluation frameworks often fail to capture the asymmetric risks inherent in medical triage. For example, standard metrics like AUROC can be highly misleading on datasets with severe class imbalance, while the standard F_1-score treats false positives and false negatives equally. To accurately assess both the discriminative power and the safety of our hybrid selective screening policy, we employ a specialized suite of clinical and probabilistic metrics.

Therefore, we evaluate: (1) Discriminative Safety via F_2-score, penalizing false negatives (misclassifying a suspect patient as non-suspect) more heavily than false positives; (2) Probabilistic Reliability via Expected Calibration Error (ECE) (Guo et al., 2017); (3) Triage Efficiency via Coverage (fraction of automated records) and True Positive Deferral Rate (TPDR), defined as the proportion of actual suspect cases referred to a physician; and (4) Uncertainty Ranking via the Area Under the Risk-Coverage curve (AURC).

Finally, to evaluate the final trinary triage decision (Clear Negative, Defer, Clear Positive) against the binary ground truth, we compute a (5) Custom Risk-Kappa score. This adaptation of Cohen's Weighted Kappa applies a strictly asymmetric penalty matrix W derived from the established preventative medicine paradigm for infectious diseases. In early HIV detection, the epidemiological cost of a missed diagnosis—delaying ART and risking transmission—far outweighs an unnecessary routine test. To formalize this, the maximum penalty of 1.0 is strictly reserved for the catastrophic failure of automating a False Negative (FN) ($W_{FN} = 1.0$). Conversely, False Positive automations ($W_{FP} = 0.5$) and the deferral of True Positive cases ($W_{def_pos} = 0.5$) incur moderate penalties, reflecting the lower clinical cost of resource utilization and the failure to automate complex cases. Safely deferring a True Negative carries the lowest penalty ($W_{def_neg} = 0.25$).

$$W = \begin{bmatrix} 0.0 & 0.5 & 1.0 \\ 0.5 & 0.25 & 0.0 \end{bmatrix}$$

A complete sensitivity analysis of these penalty weights across diverse simulated health-economic realities is provided in Appendix B.

4 Results and Discussion

To evaluate the clinical viability of the proposed selective screening framework, we benchmarked the two architectures explained in Section 3.3.2. Also, to isolate the value of our specialized architecture, we introduce a standard transformer baseline using only the base encoder (`PlanTL-GOB-ES/bsc-bio-ehr-es`).

The architectures were systematically evaluated across multiple UQ algorithms and varying risk tolerances (α). To rigorously validate statistical stability, the exhaustive evaluation—including 95% Confidence Intervals (CI) calculated via empirical bootstrap resampling ($n = 1,000$ iterations)—is provided in Appendix D.

Architecture	UQ Algorithm	Discriminative ($F_2 \uparrow$)		Calibration (ECE $\downarrow$)		Triage Efficiency & Safety			
		Binary	Clear	Binary	Clear	Coverage $\uparrow$	TPDR $\downarrow$	AURC $\downarrow$	Risk-Kappa $\uparrow$
Enc. Baseline	CV Ensemble	0.769	0.973	0.077	0.031	47.8%	50.5%	0.026	0.377
Standard MIL	MC Dropout	0.835	0.975	0.033	0.011	68.8%	34.2%	0.007	0.579
	CV Ensemble	0.823	0.979	**0.018**	**0.009**	63.5%	**29.6%**	0.007	0.546
MD-SN MIL	MC Dropout	**0.839**	0.973	0.044	0.031	70.2%	30.6%	0.011	**0.602**
	MD-SN	0.813	0.966	0.028	0.026	**70.7%**	31.9%	0.012	0.601
	CV Ensemble	0.821	**0.982**	0.022	0.021	67.7%	33.2%	**0.006**	0.576

Table 2: Discriminative performance, calibration, and triage efficiency under strict safety constraints ($\alpha = 0.01$). Metrics evaluate the forced deterministic output (Binary) and the automated subset after deferral (Clear). **Note:** Full results, including 95% bootstrap Confidence Intervals, are detailed in Appendix D.

4.1 Forced Binary Classification

Table 2 presents a head-to-head comparison under strict safety constraints ($\alpha = 0.01$). As detailed in the "Binary" columns, evaluating the cohort under a forced-classification paradigm reveals the exact vulnerability of standard NLP benchmarks. For instance, utilizing MC Dropout on the MD-SN architecture yields the highest absolute Binary F_2 score (0.839), yet this masks severely degraded calibration (ECE: 0.044) compared to the MD-SN ensemble (0.022).

A clear architectural progression is visible across the 10-model ensembles. Although the MIL architectures demonstrate a clear representational superiority over the Naive Encoder baseline ($\sim$0.05 F_2 improvement), their absolute global F_2 ($\sim$0.82) remains fundamentally unsafe for autonomous deployment. This underscores the danger of compelling a system to guess on ambiguous or OOD narratives, predictably resulting in overconfident misclassifications.

4.2 Selective Screening

Applying the Dual Veto policy at a strict safety tolerance ($\alpha = 0.01$) exposes the operational flaws of standard baselines. Hampered by poor calibration and sub-optimal uncertainty ranking, the ensembled Naive Encoder suffers a severe coverage collapse (47.8%) and a clinically unviable TPDR of 50.5%. In contrast, the ensembled MIL architectures successfully isolate highly reliable predictive domains. MD-SN MIL yields the best uncertainty ranking (AURC: 0.006 vs. 0.007 for Standard MIL) and 0.982 Clear F_2 at 67.7% coverage , effectively vetoing overconfident misclassifications.

Comparing UQ backends highlights a critical operational trade-off. While CV-Ensembles establish the empirical safety upper bound, they incur a prohibitive 10x computational overhead. Conversely, MC Dropout bypasses this footprint but struggles with probabilistic calibration in this highly imbalanced domain. Bridging this gap, the deterministic MD-SN UQ algorithm achieves a highly competitive Clear F_2 of 0.966 with superior calibration in a single pass. This efficiently isolates the safe operating domain without the stochastic latency of MC Dropout or the memory burden of ensembles, offering a viable alternative for resource-constrained deployments.

Deferral Policy	Coverage $\uparrow$	Clear F2 $\uparrow$	Risk-Kappa $\uparrow$
Aleatoric Only (MCP)	92.6% [91.0-94.0]	0.902 [0.875-0.926]	0.751 [0.713-0.789]
Epistemic Only (MCMD)	98.8% [98.2-99.3]	0.831 [0.785-0.872]	**0.788** [0.742-0.834]
Standard Uncertainty	**99.3%** [98.9-99.7]	0.824 [0.777-0.866]	**0.788** [0.741-0.833]
Dual Veto (Hybrid)	91.3% [89.8-92.7]	**0.913** [0.888-0.935]	0.755 [0.717-0.794]

Table 3: Ablation of the deferral policy components on the MD-SN model ($\alpha = 0.05$). 95% CI calculated via empirical bootstrap resampling ($n = 1,000$) are reported below each metric.

4.3 The Necessity of the Dual Veto

An ablation study (Table 3) on the MD-SN architecture ($\alpha = 0.05$) confirms that single-modality uncertainty checks are structurally insufficient. Relying exclusively on aleatoric control (MCP) successfully flags probabilistic ambiguity but remains blind to overconfident geometric anomalies. Conversely, an epistemic-only policy (MCMD) suffers a catastrophic drop in safety by unsafely automating 98.8% of the cohort. Crucially, this mirrors the failure of traditional selective classification approaches (represented here as "Standard Uncertainty" using a 90th-percentile predictive threshold), which yields a falsely optimistic 99.3% coverage but degrades the Clear F_2 to 0.824.

Although single-modality policies yield nomi-

Architecture	Risk Tol. (α)	Coverage ↑	TPDR ↓	Clear F_2 ↑	Clear ECE ↓	Risk-Kappa ↑
Standard MIL	0.01	63.5%	29.6%	0.979	**0.009**	0.546
	0.02	75.4%	19.4%	0.954	0.012	0.644
	0.05	91.1%	**7.0%**	0.902	0.015	0.747
MD-SN MIL	0.01	67.7%	33.2%	**0.982**	0.021	0.576
	0.02	77.4%	24.8%	0.969	0.028	0.665
	0.05	**91.3%**	10.2%	0.913	0.029	**0.755**

Table 4: Evaluation of the framework as an adjustable operational dial using the CV Ensemble UQ algorithm. By varying the conformal risk tolerance (α), administrators can dynamically balance system Coverage against discriminative safety (Clear F_2) and the True Positive Deferral Rate (TPDR). **Note:** Full results, including 95% bootstrap Confidence Intervals, are detailed in Appendix D.

nally higher Custom Risk-Kappa scores (0.788), their 95% CIs overlap extensively with our hybrid framework. Because of this overlap, the advantage of the dual veto on the Risk-Kappa metric is not statistically established at this current sample size. However, the hybrid policy achieves a statistically significant leap in Clear F_2 (0.913 vs. 0.824, with non-overlapping CIs). Therefore, the $\sim$8% drop in coverage represents a necessary, statistically robust trade-off to eliminate the unsafe false negatives that standard UQ backends fail to intercept.

4.4 Clinical Efficiency and the Risk Trade-off

While strict safety constraints ($\alpha = 0.01$) effectively isolate a highly reliable predictive domain, a static threshold fails to accommodate the fluctuating resource constraints of real-world healthcare systems. A clinically viable NLP triage framework must function as an adjustable operational dial, allowing hospital administrators to balance diagnostic safety with administrative workload reduction. Table 4 demonstrates the framework's scalability across broader risk thresholds ($\alpha \in \{0.01, 0.02, 0.05\}$).

Relaxing the conformal boundary on the MD-SN ensemble from a strict $\alpha = 0.01$ to a moderate $\alpha = 0.02$ allows the framework to expand its operational coverage to 77.4%, while maintaining a robust Clear F_2 of 0.969. At a more lenient $\alpha = 0.05$ setting, the system safely automates an outstanding 91.3% of patient trajectories. Even at this high coverage level, the framework preserves a Clear F_2 of 0.913 and a low TPDR of 10.2%.

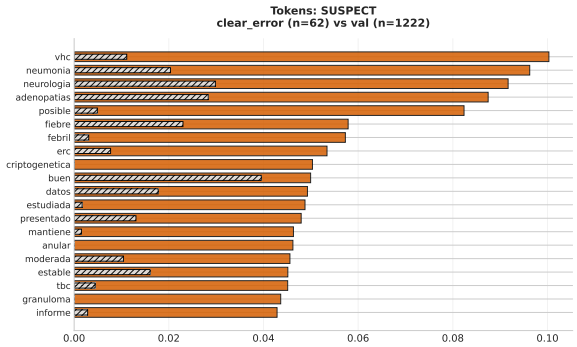
Furthermore, evaluating the Standard MIL architecture reinforces the system's architecture-agnostic flexibility. At $\alpha = 0.05$, the Standard MIL provides comparable coverage (91.1%) with superior calibration (ECE: 0.015) and an exceptionally low TPDR (7.0%), albeit with slightly lower discriminative safety. This performance parity proves that the dual-veto framework successfully rescues uncalibrated baselines and provides administrators with viable operational alternatives: prioritizing maximum safety (MD-SN MIL) or minimizing expert workload with superior calibration (Standard MIL).

4.5 Interpretability of the Triage Decisions

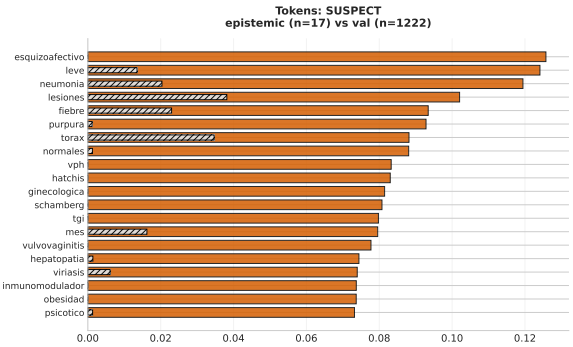
To visualize the underlying clinical reasoning and mitigate the black-box nature of deep neural networks, we analyzed feature attributions at the $\alpha = 0.05$ threshold. Crucially, to accurately capture representational shifts, we explicitly compared the marginal feature attributions of the various test triage cohorts against the mean feature attributions of the validation cohort, which is our established, in-distribution baseline. While the absolute number of instances in these specific triage subsets is naturally small (e.g., 62 for automated FP, 17 for epistemic suspicion deferrals), they encompass the entirety of these edge cases within the test set[2]. A detailed methodological explanation of our custom hierarchical attribution system, alongside additional visual examples, is provided in Appendix C.

When analyzing the system's "Clear Errors", specifically FP where the model predicted clinical suspicion but no serology was documented, the feature attributions reveal strong clinical robustness when measured against this validation baseline. As shown in Figure 2a, these predictions are heavily driven by explicit indicator conditions mandating screening under EuroTEST guidelines, such as *vhc* (Hepatitis C) and *tbc* (Tuberculosis). This suggests the model successfully anchors on guideline-concordant phenotypic risk; the absence of a test highlights the noisy, partially observable

[2]Aleatoric deferrals are excluded from this analysis, as probabilistic ambiguity is driven by the joint conflict of typical features, which marginal aggregations cannot adequately capture.



(a) Automated False Positives ("Clear Errors")



(b) Epistemic Deferrals (Suspicion)

Figure 2: Feature attribution comparison for complex triage edge cases at $\alpha = 0.05$. Solid colored bars represent the mean token-level importance for the positive class on the designated test cohort. The underlying light grey bars with black diagonal hatching denote the validation baseline.

nature of EHR documentation rather than a representational failure.

Conversely, instances flagged by the epistemic geometric veto demonstrate severe representational collapse. As illustrated in Figure 2b, predictions for these deferred cases are hijacked by highly specific, OOD artifacts, such as psychiatric conditions (*esquizoafectivo-schizoaffective-*, *psicótico-psychotic-*) or unrelated localized infections (*vulvovaginitis*), which registered near-zero predictive importance during validation. By calculating the Mahalanobis distance, the framework successfully recognizes this latent space distortion and safely vetoes the prediction before it can confidently misclassify the atypical presentation.

5 Conclusion

Translating biomedical NLP into clinical workflows is fundamentally limited by standard evaluation paradigms that force deterministic decisions on ambiguous narratives. This inherently produces overconfident misclassifications that directly conflict with medical safety requirements. To bridge this gap, we introduced and evaluated a risk-aware, hybrid selective screening framework for early HIV suspicion identification in a highly imbalanced Spanish EHR cohort.

Our empirical evaluations demonstrate that standard UQ methodologies conflate distinct predictive risks, forcing baseline models into severe coverage collapse when subjected to strict reliability constraints. In contrast, explicitly decoupling aleatoric uncertainty (via MCP) and epistemic uncertainty (via a MCMD veto) allows the evaluation framework to isolate a highly trustworthy operational domain. By functioning as an adjustable operational dial, the system enables healthcare administrators to dynamically balance diagnostic safety with administrative workload. Under strict safety constraints ($\alpha = 0.01$), the dual-veto framework maximizes discriminative safety, achieving a high Clear F_2-score of 0.982 across 67.7% of the cohort. When relaxed to accommodate high-throughput workloads ($\alpha = 0.05$), the framework safely automates a 91.3% of patient trajectories while preserving a robust 0.913 Clear F_2 and limiting true positive deferrals to just 10.2%. Beyond quantitative safety metrics, our hierarchical feature attribution analysis confirms the clinical robustness of the triage logic: automated false positives reliably anchor on guideline-concordant phenotypic risks, while the epistemic veto proactively intercepts semantic artifacts and atypical presentations.

Ultimately, this study shows that optimizing biomedical NLP systems purely for raw classification accuracy creates a dangerous illusion of clinical readiness. By ensuring that highly ambiguous patient profiles and geometric anomalies are proactively flagged for expert review, our framework proves that rigorous, decoupled uncertainty quantification is a fundamental prerequisite for translating NLP into responsible clinical practice.

Building upon these findings, future work will focus on adapting this risk-aware triage paradigm to diverse clinical use cases where the asymmetric costs of errors differ from the preventative infectious disease model.

To support methodological transparency, the complete UQ implementation pipeline and inference code is publicly available on GitHub[3].

[3] https://github.com/romorale/mil_uq_public

Limitations

While the framework bridges theoretical UQ and operational clinical safety, key methodological limitations remain. First, while the sensitivity analysis in Appendix B validates the robustness of the Custom Risk-Kappa under various simulated health-economic realities, it remains an experimental metric strictly tailored to the specific risk profile of early infectious disease screening. In this context, the severe epidemiological consequence of a missed diagnosis heavily outweighs the impact of an unnecessary serological test. However, applying this metric to domains with high physical intervention risks (e.g., invasive surgical biopsies in oncology) requires fundamental recalibration and formal health-economic validation.

Second, the static geometric thresholds τ_{dist} make the epistemic veto vulnerable to temporal or demographic dataset shifts prior to scheduled retraining. Third, although generative Large Language Models (LLMs) dominate current NLP benchmarks, extracting reliable, well-calibrated uncertainty from autoregressive generation remains an unsolved clinical challenge. Therefore, we explicitly prioritized a deterministic, spectrally normalized encoder to guarantee the geometric vetoes required for strict triage safety.

Finally, empirical validation is constrained to a single medical center. Given that HIV suspicion relies on partially observable behavioral indicators, cross-institutional validation is necessary to confirm the framework's broader generalizability.

Ethical Considerations

The use of data for this study has been implemented in full compliance with strict adherence to the European General Data Protection Regulation (GDPR) and the Spanish Organic Law 3/2018 (LOPDGDD) on the Protection of Personal Data and Guarantee of Digital Rights. Due to strict institutional ethics board regulations privacy constraints regarding highly sensitive clinical narratives, the raw electronic health records cannot be made publicly accessible.

Acknowledgements

This work was supported by the projects EDHER-MED (PID2022-136522OB-C21; MCIN/AEI/10.13039/501100011- 033/FEDER, UE), funded by the Spanish State Research Agency and the Spanish Ministry of Science, Innovation and Universities.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76:243–297.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. 2021. Mondrian conformal predictive distributions. In Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications, volume 152 of Proceedings of Machine Learning Research, pages 24–38. PMLR.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 193–199, Stroudsburg, PA, USA. Association for Computational Linguistics.

EuroTEST. 2016. HIV Indicator Conditions: Guidance for Implementing HIV Testing in Adults in Health Care Settings.

Daniel J. Feller, Jason Zucker, Michael T. Yin, Peter Gordon, and Noémie Elhadad. 2018. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. JAIDS Journal of Acquired Immune Deficiency Syndromes, 77(2):160–166.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning, volume 48 of ICML'16, page 1050–1059. JMLR.org.

Juan Jose Garcia, Rebecca Kitzmiller, Ashok Krishnamurthy, and Jessica K. Zègre-Hemsey. 2026. Selective classification with machine learning uncertainty estimates improves ACS prediction: a retrospective study in the prehospital setting. Scientific Reports, 16(1):913.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of ICML'17, page 1321–1330. JMLR.org.

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 2127–2136. PMLR.

Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1):195.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: regularized dropout for neural networks. In Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2999–3007.

Sravanthi Machcha, Sushrita Yerra, Sharmin Sultana, Hong Yu, and Zonghai Yao. 2025. Do large language models know when not to answer in medical QA? In Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 27–35, Suzhou, China. Association for Computational Linguistics.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In Proceedings of the International Conference on Learning Representations (ICLR 21).

Rodrigo Morales-Sánchez, Soto Montalvo, Adrián Riaño, Raquel Martínez, and María Velasco. 2024. Early diagnosis of HIV cases by means of text mining and machine learning models on clinical notes. Computers in Biology and Medicine, 179:108830.

Francis Ngema, Albert Whata, Micheal O Olusanya, and Siyabonga Mhlongo. 2026. Explainable AI–Driven Comparative Analysis of Machine Learning Models for Predicting HIV Viral Nonsuppression in Ugandan Patients: Retrospective Cross-Sectional Study. JMIR AI, 5:e68196.

Alina Peluso, Ioana Danciu, Hong-Jun Yoon, Jamaludin Mohd Yusof, Tanmoy Bhattacharya, Adam Spannaus, Noah Schaefferkoetter, Eric B. Durbin, Xiao-Cheng Wu, Antoinette Stroup, Jennifer Doherty, Stephen Schwartz, Charles Wiggins, Linda Coyle, Lynne Penberthy, Georgia D. Tourassi, and Shang Gao. 2024. Deep learning uncertainty quantification for clinical text classification. Journal of Biomedical Informatics, 149:104576.

Ashok Kumar Sah, Rabab H. Elshaikh, Manar G. Shalabi, Anass M. Abbas, Pranav Kumar Prabhakar, Asaad M. A. Babker, Ranjay Kumar Choudhary, Vikash Gaur, Ajab Singh Choudhary, and Shagun Agarwal. 2025. Role of Artificial Intelligence and Personalized Medicine in Enhancing HIV Management and Treatment Outcomes. Life, 15(5):745.

Mame Diarra Toure and David A. Stephens. 2026. Not just how much, but where: Decomposing epistemic uncertainty into per-class contributions. arXiv:2602.21160.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty Estimation of Transformer Predictions for Misclassification Detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8237–8252, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angelo Ziletti and Leonardo D'Ambrosi. 2026. Disentangling ambiguity from instability in large language models: A clinical text-to-sql case study. arXiv:2602.12015.

A Hyperparameter Sensitivity Analysis

To ensure the robustness of the epistemic veto, we evaluated the framework's triage efficiency across varying inertia convergence limits (0.01, 0.02, 0.05, 0.1) and Mahalanobis distance percentiles (90th, 95th, 99th, 99.5th, 99.9th) on the test set, calibrated over the validation set. As shown in Table A1, the chosen configuration (0.05, 99th) represents the empirical global optimum, maximizing the clinical safety metric (Custom Risk-Kappa) while maintaining a low TPDR.

B Custom Risk-Kappa ablation and the impact of model coverage

To evaluate the behavioral differences between the Standard MIL and the MD-SN MIL architectures under varying operational constraints, we conducted an ablation of the Custom Risk-Kappa metric across five simulated health-economic realities.

Table B1 demonstrates how these two architectures perform depending on the institutional penalties assigned to False Positives (FP), Deferred True Negatives (Def_TN), and Deferred True Positives (Def_TP). A missed diagnosis (False Negative) is consistently assigned the maximum catastrophic penalty of 1.0 across all scenarios.

Across all simulated environments, the MDSN MIL consistently outperforms the Standard MIL

Inertia (Δ)	Dist. Percentile	Coverage ↑	Clear F_2 ↑	TPDR ↓	Risk-Kappa ↑
Strict Extreme					
0.01	90th	84.3%	0.946	19.7%	0.721
Smooth Degradation					
0.05	90th	84.7%	0.945	19.7%	0.724
0.05	95th	87.5%	0.935	16.1%	0.746
Chosen Optimum					
0.05	**99th**	**91.3%**	**0.913**	**10.1%**	**0.755**
Loose Extreme					
0.10	99.9th	92.4%	0.904	8.7%	0.752

Table A1: Hyperparameter Sensitivity Analysis across varying inertia convergence limits (0.01, 0.02, 0.05, 0.1) and Mahalanobis distance percentiles (90th, 95th, 99th, 99.5th, 99.9th) on the test set.

Health-Economic Reality	Penalty Weights				Standard MIL (63.5% Cov.)	MDSN MIL (67.7% Cov.)
	FN	FP	Def_TN	Def_TP		
Zero-Cost Deferral *(Maximum Caution)*	1.0	1.0	0.0	0.0	0.948	0.960
Extreme Epidemiological *(Resource-Rich)*	1.0	0.25	0.1	0.25	0.695	0.723
Default Baseline *(Preventative Medicine)*	**1.0**	**0.5**	**0.25**	**0.5**	**0.546**	**0.576**
Symmetric Control *(Classification Baseline)*	1.0	1.0	0.5	0.5	0.475	0.513
High Admin Burden *(Resource-Constrained)*	1.0	0.75	0.5	0.75	0.432	0.463

Table B1: Custom Risk-Kappa Scores Across Simulated Health-Economic Realities.

baseline. In a "Zero-Cost Deferral" scenario representing maximum caution (where manual hospital review incurs no penalty), the MDSN MIL achieves an exceptional Risk-Kappa of 0.960 compared to the Standard MIL's 0.948. This highlights that when an institution has unlimited bandwidth to review deferrals, the MDSN's rigid safety mechanisms provide near-perfect predictive accuracy on its covered subset.

Conversely, when simulating a resource-constrained environment ("High Administrative Burden"), penalties for false positives (0.75) and deferrals (0.5 to 0.75) are heavily increased. Even under these strict conditions, the MDSN MIL (0.463) maintains a clear advantage over the Standard MIL (0.432). This consistent superiority across both extreme epidemiological and highly constrained administrative realities proves that the MD-SN architecture's superior calibration and epistemic veto mechanism successfully isolate ambiguities without overwhelming institutional resources.

C Feature Attribution Analysis

C.1 Hierarchical Feature Attribution Methodology

A common limitation in interpreting Multiple Instance Learning (MIL) architectures is the difficulty of projecting local token importances into the global document representation (z_{doc}). To generate the feature attributions, we implemented a custom hierarchical attention-occlusion explainer, computed in four stages:

1. **Local Token Importance:** Token-level importance is extracted from the base transformer's last-layer attention weights (combining [CLS] and self-attention) and aggregated into continuous word-level scores.

2. **Global MIL Aggregation:** To project local scores to the patient level, each word's score is multiplied by the normalized MIL gated-attention weight (w_c) assigned to its respective chunk. The global importance is the sum of its MIL-weighted occurrences:

$$I(\text{word}) = \sum_{c \in \text{chunks}} \text{Attn}(\text{word}_c) \cdot w_c$$

3. **Directional Impact (Occlusion):** Because attention only provides magnitude, we determine polarity via a leave-one-out occlusion test. Masking the target word across all chunks shifts the decision boundary; this logit margin (Δ) determines whether the token drives the prediction toward Suspicion or Non-Suspicion.

4. **Cohort TF-IDF Aggregation:** To prevent highly-attended outlier documents from skewing the cohort average, patient-level importances are aggregated using a modified TF-IDF weighting. Term Frequency (intra-patient importance) is square-root scaled to dampen extremes, while the IDF penalty upweights consistently important terms across the cohort. Finally, these scores are averaged across the target subset to yield the stable attributions displayed in the figures.

C.2 Supplementary Triage Visualizations

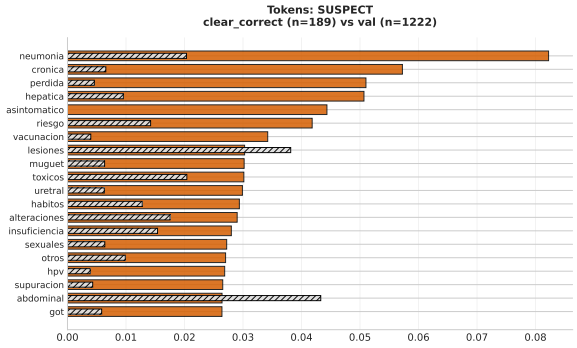
This section provides supplementary visualizations to confirm the model's expected baseline behavior and further demonstrate the necessity of the epistemic veto.

Figure C1 illustrates the feature attributions for the remaining triage cohorts. Plot C1a confirms the system's reliability on standard, safely automated true positive cases ("Clear Correct"). In these instances, the model's predictions rely entirely on established, in-distribution clinical indicators (such as *neumonía-pneumonia-* and *muguet*), confirming that its automated decisions are grounded in relevant epidemiological phenotypes

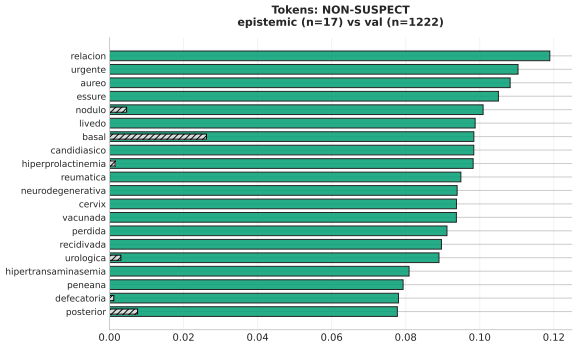
Conversely, Plot C1b demonstrates the critical role of the geometric veto even for cases where the model predicts the negative class (Non-Suspicion). In these deferred cases, the model is distracted by highly specific, OOD medical artifacts that distort the latent space. A prime example is the heavy attribution given to *essure* (a specific gynecological implant). While such terms may spuriously co-occur with routine administrative testing in the training data, they are fundamentally unrelated to the targeted clinical construct. The MCMD successfully recognizes this geometric anomaly and intercepts the prediction, preventing the model from making an overconfident decision based on irrelevant clinical noise.

D Exhaustive Evaluation Metrics

Table D1 provides the comprehensive triage metrics across all tested architectures, Uncertainty Quantification algorithms, and risk tolerances. All metrics include 95% Confidence Intervals (CI) derived via empirical bootstrap resampling ($n = 1,000$).



(a) Automated True Positives ("Clear Correct")



(b) Epistemic Deferrals (Non-Suspicion)

Figure C1: Feature attribution comparison for complex triage edge cases at $\alpha = 0.05$. Solid colored bars (orange for the positive class, green for the negative class) represent the mean token-level importance for the designated test cohort. The underlying light grey bars with black diagonal hatching denote the validation baseline.

Architecture	**UQ Algorithm** (α)	**Discriminative** ($F_2 \uparrow$)		**Calibration (ECE**$\downarrow$**)**		**Triage Efficiency & Safety**			
		Binary	**Clear**	**Binary**	**Clear**	**Coverage** $\uparrow$	**TPDR** $\downarrow$	**AURC** $\downarrow$	**Risk-Kappa** $\uparrow$
Enc. Baseline	CV Ens (0.01)	0.769 [0.722-0.811]	0.973 [0.958-0.985]	0.077 [0.063-0.092]	0.031 [0.021-0.042]	47.8% [45.2-50.3]	50.5% [43.9-57.3]	0.026 [0.020-0.032]	0.377 [0.341-0.410]
Standard MIL	MC Dropout (0.01)	0.835 [0.793-0.874]	0.975 [0.954-0.991]	0.033 [0.024-0.043]	0.011 [0.006-0.016]	68.8% [66.4-71.2]	34.2% [28.2-40.5]	0.007 [0.005-0.009]	0.579 [0.543-0.616]
	CV Ens (0.01)	0.823 [0.776-0.864]	0.979 [0.965-0.991]	**0.018** [0.010-0.025]	**0.009** [0.004-0.015]	63.5% [60.8-66.1]	**29.6%** [23.6-35.7]	0.007 [0.005-0.010]	0.546 [0.507-0.582]
	CV Ens (0.02)	0.823 [0.776-0.864]	0.954 [0.931-0.973]	0.018 [0.010-0.025]	0.012 [0.006-0.019]	75.4% [73.0-77.6]	19.4% [14.4-24.6]	0.007 [0.005-0.010]	0.644 [0.606-0.679]
	CV Ens (0.05)	0.823 [0.776-0.864]	0.902 [0.872-0.929]	0.018 [0.010-0.025]	0.015 [0.008-0.024]	91.1% [89.7-92.6]	7.0% [3.9-10.5]	0.007 [0.005-0.010]	0.747 [0.705-0.783]
MD-SN MIL	MC Dropout (0.01)	**0.839** [0.797-0.877]	0.973 [0.955-0.988]	0.044 [0.034-0.054]	0.031 [0.024-0.039]	70.2% [67.8-72.5]	30.6% [24.7-37.0]	0.011 [0.006-0.016]	**0.602** [0.566-0.639]
	MD-SN (0.01)	0.813 [0.767-0.856]	0.966 [0.945-0.983]	0.028 [0.019-0.038]	0.026 [0.017-0.034]	**70.7%** [68.4-73.0]	31.9% [25.9-38.2]	0.012 [0.006-0.020]	0.601 [0.565-0.638]
	CV Ens (0.01)	0.821 [0.774-0.864]	**0.982** [0.979-0.995]	0.022 [0.013-0.031]	0.021 [0.015-0.027]	67.7% [65.3-70.4]	33.2% [27.1-40.2]	**0.006** [0.004-0.008]	0.576 [0.539-0.613]
	CV Ens (0.02)	0.821 [0.774-0.864]	0.969 [0.954-0.983]	0.022 [0.013-0.031]	0.028 [0.021-0.036]	77.4% [75.2-79.8]	24.8% [19.4-30.7]	0.006 [0.004-0.008]	0.665 [0.629-0.701]
	CV Ens (0.05)	0.821 [0.774-0.864]	0.913 [0.888-0.935]	0.022 [0.013-0.031]	0.029 [0.021-0.038]	91.3% [89.8-92.7]	10.2% [6.2-14.2]	0.006 [0.004-0.008]	0.755 [0.717-0.794]

Table D1: Discriminative performance, calibration, and triage efficiency across models and Uncertainty Quantification (UQ) backends. Metrics evaluate the forced deterministic output (Binary) and the automated subset after deferral (Clear). Triage safety is further assessed via Coverage, True Positive Deferral Rate (TPDR), Area Under the Risk-Coverage curve (AURC), and Custom Risk-Kappa. 95% Confidence Intervals (CI) calculated via empirical bootstrap resampling ($n = 1,000$) are reported below each metric. The best performance per metric in the strict safety tier ($\alpha = 0.01$) is highlighted in bold.

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

Julien Sylvestre
BXCT Partners Ltd
js@bxct.co.uk

Abstract

SciFact is a widely-used benchmark for scientific claim verification (645 citations, included in the BEIR evaluation suite). We present, to our knowledge, the first systematic annotation audit of its development and training sets, combining automated screening with a small language model ($0.11 in API fees) and exhaustive manual verification against source publications. We identify 11 gold-label errors in the development set (5.3%, 95% CI 2.7–9.2%, of 209 audited claim–document pairs) and 13 in the training set (2.3%, 95% CI 1.2–3.9%, of 564 audited pairs). The dev errors exhibit a directional asymmetry—9 of 11 mislabel a claim as SUPPORT (one-sided binomial p=0.033, two-sided p=0.065)—and fall into four recurring types. Correcting the dev labels raises binary macro-F1 by 1.7–3.8 points across GPT-5.4 (mini, nano) and Claude Haiku 4.5; gains are larger in 3-way evaluation when mislabeled evidence is recast as NEI (e.g., +9.2 with Haiku 4.5). The binary range is comparable in magnitude to inter-system margins on the SciFact leaderboard. A simple claim-only probe with Haiku 4.5 does not support label memorization as the main explanation for these gains. We release corrected annotations and a blind annotator packet, and recommend that benchmark users prefer the corrected release going forward.

1 Introduction

Benchmark quality is foundational to NLP progress: if gold labels contain systematic errors, model comparisons become unreliable and optimization targets become misaligned. This concern is especially acute for small, expert-annotated datasets where a handful of label errors can shift rankings by several F1 points.

SciFact (Wadden et al., 2020), introduced at EMNLP 2020 and subsequently incorporated into the BEIR retrieval evaluation suite (Thakur et al., 2021), is the standard benchmark for scientific claim verification (645 citations on Semantic Scholar as of April 2026). The task requires systems to retrieve relevant abstracts and classify scientific claims as SUPPORT, CONTRADICT, or NOT ENOUGH INFO (NEI) given evidence sentences from PubMed abstracts. The dataset contains 300 dev claims, 809 training claims, and 300 hidden test claims; of these, 209 dev and 564 training claim–document pairs have gold evidence annotations that we audit here. Gold labels were assigned by domain experts (reported Cohen's $\kappa = 0.75$). Despite wide adoption—over 600 citations and use in recent work at EMNLP, SIGIR, and JAMIA Open—to our knowledge no prior study has audited the quality of SciFact's gold annotations. A review of the dataset's GitHub repository, Hugging Face page, and citing literature (conducted March 2026) reveals zero reported label corrections, no errata, and no corrected release in six years of use. This is notable given that the annotated pairs are small enough to verify manually in a few hours and that the original inter-annotator agreement ($\kappa = 0.75$) leaves room for disagreement.

We contribute: (1) identification of 11 gold label errors in the dev set (5.3%, 95% CI 2.7–9.2%) and 13 in the training set (2.3%, 95% CI 1.2–3.9%), showing errors contaminate both model training and evaluation; (2) a taxonomy of error types and analysis of directional asymmetry, with statistical framing; (3) cross-model verification of the F1-impact and a contamination probe using a non-GPT model family (Claude Haiku 4.5); and (4) an efficient, replicable audit methodology costing $0.11 in API fees for the automated screening stage.

2 Related Work

Annotation errors in benchmarks. Northcutt et al. (2021) found 3–6% label errors across 10 pop-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 97–103
July 3-4, 2026 ©2026 Association for Computational Linguistics

ular benchmarks (including ImageNet and MNIST) and showed that correcting these errors changes model rankings. More recently, Vendrow et al. (2025) created "platinum" corrected versions of 15 benchmarks and found that over 50% of model errors on the originals were attributable to label noise—GSM8K alone had ∼5% problematic questions. Nahum et al. (2025) used LLM ensembles to detect 6–21% error rates across factual consistency benchmarks, showing that many reported LLM "mistakes" are correct responses to mislabeled examples. In NER, Rucker and Akbik (2023) re-annotated CoNLL-03 and found 7% of labels were incorrect; nearly half of state-of-the-art model "errors" were actually correct predictions penalized by noisy ground truth.

Annotation artifacts and disagreement. Gururangan et al. (2018) identified annotation artifacts in SNLI; Pavlick and Kwiatkowski (2019) and Nie et al. (2020) showed many NLI items lack a single correct label. Weber-Genzel et al. (2024) separated genuine label variation from error in MNLI. Swayamdipta et al. (2020) found that hard-to-learn instances often correspond to labeling errors, and Klie et al. (2023) surveyed 18 error detection methods, finding the field lacks standardized evaluation.

SciFact. Wadden et al. (2022b) showed 15+ F1 drops when expanding SciFact to open-domain retrieval, and Košprdić et al. (2024) found fine-tuned DeBERTa outperformed GPT-4 zero-shot by ∼7%. Despite this active research community, to our knowledge no prior work has audited SciFact's gold labels. The dataset's GitHub repository contains 27 issues—none concerning annotation quality—and neither the original authors nor any of the 600+ citing papers have published corrections or errata.

3 Method

3.1 Stage 1: Automated LLM Screening

We processed all 209 dev claim–document pairs using GPT-5.4-mini (temperature 0.3) with a structured audit prompt. For each pair, the model received the claim text, the gold label, the gold rationale sentences, and the full abstract. The model returned structured JSON: assessment (AGREE/DISAGREE/DEBATABLE), reasoning, error type (from a predefined taxonomy), and suggested correct label.

The automated screening flagged 57 claims (27.3%): 43 as DISAGREE and 14 as DEBATABLE. This stage cost $0.11 in API fees (approximately 440K tokens at March 2026 pricing; cached-input rates applied).

3.2 Stage 2: Manual Expert Verification

The automated screening has a high false-positive rate: the LLM frequently confuses claim direction with evidence direction (Section 4.3). Each of the 57 mini-flagged disagreements was therefore adjudicated by a single annotator working interactively with a *frontier* LLM (GPT-5.4) in chat mode: the frontier model surfaced its own reading of the claim and evidence, the annotator probed and challenged it, and the annotator made the final verdict. This human-in-the-loop protocol provides an additional model-based second opinion on every disputed pair while keeping final adjudication with the human; the screening model (mini) and the verification model (frontier) are different scales of the same family, and the human breaks ties. The 152 non-flagged claims were exhaustively reviewed by the same annotator without the chat-mode loop, since the screen had already implicitly endorsed the gold label. Corrected labels are determined solely by the assigned evidence paper; later literature is consulted only for context (see footnotes in §4.1).

Of the 57 flagged claims, 8 were ultimately confirmed as annotation errors, 8 remained debatable but defensible, and 41 were false alarms (after a second-pass review on the borderline cases: 2 claims initially classified as debatable were upgraded to confirmed errors and are listed in the appendix). Of the 152 non-flagged claims, exhaustive review identified 3 additional errors that the automated screening missed entirely (rating them AGREE), bringing the total to 11.

3.3 Source-Paper Verification

All 11 dev errors were verified against the original source publications (retrieved via PubMed), confirming that each correction is consistent with the assigned evidence paper's own findings. This verification is independent of the benchmark labels, not of the annotator: a second independent rater was not used (see Limitations).

4 Results

4.1 Confirmed Annotation Errors

Table 1 presents the 11 confirmed errors, falling into four types: **(1) Label reversal** (4 cases): gold

label directly contradicts the evidence. ID 593 claims incidence "decreased by 10%" but the rationale states "did not change over time." ID 879 is labeled CONTRADICT, yet the abstract concludes "the large majority of lincRNAs do not function through encoded proteins" (sentence 6), agreeing with the claim.[1] ID 1385's rationale states "enhancing cSMAC formation *reduced* stimulatory capacity," the opposite of the claim's "enhances," yet gold says SUPPORT.[2] ID 808's rationale is a methodology sentence; the abstract's actual finding—"does not rely on cis-acting sequences" (sentence 3)—directly contradicts the claim, yet gold says SUPPORT. **(2) Outcome mismatch** (3 cases): claim and evidence address different clinical outcomes. ID 343 claims "bleeding events" but the rationale reports cardiovascular mortality, MI, stroke, and CHF—bleeding is never mentioned. ID 1368 claims effect on "term of delivery" (preterm birth) but the rationale reports gestational diabetes, pre-eclampsia, and SGA—preterm birth is absent from the abstract. ID 770 claims "reduced efficacy and lower quality of life" for single-agent fluoropyrimidines vs. oxaliplatin, but PFS was non-significant (p=0.07) and the trial's QoL comparison was between two fluoropyrimidines (fluorouracil vs. capecitabine), not single-agent vs. combination. **(3) Entity mismatch** (3 cases): claim and evidence refer to different biological entities. ID 1216 claims cleavage in "human beta cells" but the paper reports exclusively on mouse Tmem27 (Tcf1$^{-/-}$ mice, transgenic mice).[3] ID 1274 attributes T6SS mechanisms to *E. coli* but the paper demonstrates them in *V. cholerae* and *A. baylyi*—organisms from different taxonomic orders. ID 847 claims "new drugs" fail to penetrate necrotic lesions; the rationale cites rifampicin (a 1960s drug, not new) accumulating in caseum, while the same abstract reports that moxifloxacin (a newer drug) "does not diffuse well." **(4) Overgeneralization** (1 case): ID 208 claims "CHEK2 is not associated with breast cancer," but the paper's abstract states "Previous investigation has established a role for the CHEK2 gene in breast cancer aetiology" (sentence 1) and reports OR 2.26 for the rare 1100delC mutation (sentence 8).

ID	Gold→Corr.	Type	Key issue
593	S→C	Reversal	"decreased 10%" vs "did not change"
879	C→S	Reversal	Abstract [6] agrees with claim
1385	S→C	Reversal	"enhances" vs "reduced capacity"
808	S→C	Reversal	Abstract [3] contradicts claim
343	S→N	Outcome	CV events, not bleeding
1368	S→N	Outcome	Diabetes/SGA, not preterm birth
770	S→N	Outcome	PFS p=0.07; QoL compared wrong arms
1216	S→N	Entity	Mouse study, claim says "human"
1274	S→N	Entity	*V. cholerae* study, claim says *E. coli*
847	C→N	Entity	Old drug in rationale; claim says "new"
208	S→C	Overgen.	Abstract [1,8] contradict claim

Table 1: The 11 confirmed gold label errors (n=209 dev claim–document pairs). S=SUPPORT, C=CONTRADICT, N=NEI. 9/11 are mislabeled SUPPORT. One additional per-document error (ID 597, mortality≠incidence) is not shown.

4.2 Impact on System Evaluation

We ran two zero-shot configurations on all 209 dev claims (oracle setting) and computed macro F1 before and after correcting the 11 errors. GPT-5.4-mini gained +3.8 F1 (84.9→88.7); GPT-5.4-nano gained +1.7 (79.5→81.2). For context, the top systems on the SciFact test-set leaderboard are separated by just 1.25 F1 points (MultiVerS at 72.5 vs. ARSJoint at 71.2; Wadden et al., 2022a; Zhang et al., 2021), and dev-set gaps among top models range from 0.7 to 2.7 F1 (Košprdić et al., 2024). The correction-induced shifts of 1.7–3.8 F1 are of the same order as these gaps, introducing noise that can plausibly perturb leaderboard ordering—though our zero-shot estimates do not establish reordering of fine-tuned systems, whose predictions we do not have. The stronger model benefits more because the mislabeled claims were among its "errors"—it correctly rejected claims that the

[1] Subsequent studies have identified functional micropeptides encoded by some lncRNAs, complicating the blanket claim. However, the *cited paper's own conclusion* supports the claim; the gold label is wrong relative to the assigned evidence regardless of later findings.

[2] A follow-up by the same group (Cemerski et al., 2008) refined this finding, reporting that cSMAC can enhance signaling from weak agonists under certain conditions. Nevertheless, the *assigned evidence paper* unambiguously states "reduced," making the SUPPORT label incorrect for this claim–evidence pair.

[3] Later work confirmed TMEM27 cleavage in human beta cells (Esterházy et al., 2011). The claim is likely true, but the *assigned evidence* is a mouse-only study and cannot support the species-specific claim.

gold standard incorrectly accepted.

Cross-model verification. To address the concern that the F1 gain might reflect GPT-family preferences leaking into our corrections, we reran the same evaluation with Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) using identical inputs (claim, paper title, full abstract, gold rationale), structured tool-use output, and prompt caching. On the original labels, Haiku reaches a binary macro F1 of 92.77; applying the corrections raises it to 95.75—a $+2.98$ F1 gain, directionally consistent with GPT-5.4-mini's $+3.8$. The 3-way (S/C/NEI) macro F1 gain is larger ($61.85 \rightarrow 71.09$, $+9.24$), reflecting the 6 pairs whose evidence licenses NEI rather than a binary verdict. Across both model families and both binary and 3-way evaluation, every configuration gains F1 from the corrections; the effect is not confined to the GPT family in our tests.

Contamination probe. We then asked Haiku 4.5 to classify the same 209 dev claims given *only* the claim text—no abstract, no rationale, no title. Haiku reaches 70.81% accuracy, only 4.78 points above the 66.03% majority-class baseline (SUPPORT), with 18% of predictions being NEI. A model that had memorized SciFact's binary gold labels would be expected to commit to a side; the small gap above majority class is consistent with claim-plausibility reasoning rather than label memorization. This argues against dataset contamination as the source of the F1 gain.

4.3 Automated Auditor Performance

Of 57 flagged claims, 8 were errors (14% precision). Most false alarms were *direction confusion*: the auditor flagged CONTRADICT labels as wrong when the claim reversed the evidence, not recognizing that CONTRADICT is correct *because* the claim reverses the evidence. Crucially, three errors (IDs 1385, 770, 1216) were missed entirely—the LLM rated them AGREE despite clear mismatches. ID 1385 is a direct claim–evidence contradiction; ID 770 requires checking statistical significance and QoL directionality; ID 1216 requires noticing a species mismatch (mouse vs. human). The auditor's recall was 8/11 (73%); automated screening alone is insufficient.

4.4 Per-Document Error

One additional per-document error (not counted above): ID 597's third evidence document discusses cervical cancer *mortality* but is labeled SUPPORT for an *incidence* claim.

4.5 Training Set Audit

To assess whether errors also contaminate model training, we manually audited all 564 claim–document pairs in the training set. We identified 13 errors (2.3%, 95% CI 1.2–3.9%), exhibiting the same error types as the dev set; the dev and train CIs overlap, so we do not claim that dev is strictly noisier than train. Seven are clear label reversals: e.g., ID 164 labels SUPPORT for "bariatric surgery increases colorectal cancer" when the rationale states "no association was detected between bariatric surgery and … cancer"; ID 933 labels SUPPORT for "increased morphine use" when the evidence reports significantly *less* morphine use ($P{=}.03$). Three are entity mismatches: e.g., ID 621 claims "Individuals with Alzheimers" but the study explicitly recruited adults who "did not meet criteria for dementia"; ID 657 attributes a function to signal peptide peptidase but the evidence is about RHBDL4, a different enzyme. One is a causal misread: ID 416 claims APOE4 carriers "have longer lifetime exposure to estrogen due to an increased reproductive period," but the evidence only shows that APOE4 modifies the effect of reproductive period length on dementia risk (effect modification), not that APOE4 causes longer reproductive periods.[4] Two involve modality or numerical mismatches between claim and evidence (ID 1145 treats a microsimulation projection as observed data; ID 1297 claims 30 million cases when the evidence reports $\sim$15 million hospitalizations). All 13 training errors with claim IDs, corrected labels, rationale indices, and one-sentence justifications are provided in the supplementary materials.[5] The lower error rate in training (2.3% vs. 5.3% in dev) may reflect differences in annotation workflow, but the key finding is that errors are present in *both* splits: models fine-tuned on SciFact are optimized toward wrong labels during training, tuned against wrong labels during development, and—if the hidden test set has similar rates—evaluated against wrong labels at submission. This means label noise

[4] A 1.8-year delay in menopause among APOE3/4 carriers has been reported in one Chinese cohort (Meng et al., 2012), but findings are mixed across populations, and the *assigned evidence paper* reports only effect modification, not a causal relationship.

[5] Full training-set error table in supplementary materials and at `https://github.com/Kefez/scifact-audit-bionlp2026`.

can affect every stage of the SciFact pipeline, from retrieval model training (learning which documents to retrieve for a claim) to inference model training (learning which label to assign) to final evaluation.

5 Discussion

Our 5.3% dev error rate (95% CI 2.7–9.2%) falls within the 3–7% range reported for other benchmarks (Northcutt et al., 2021; Rucker and Akbik, 2023; Vendrow et al., 2025), suggesting that expert-annotated scientific datasets are not immune to label noise. The directional asymmetry (9/11 mislabeled SUPPORT) is, on its own sample size, suggestive but not conclusively significant: an exact two-sided binomial test against p=0.5 yields p=0.065 (one-sided p=0.033). The 95% Clopper–Pearson CI on the SUPPORT-error proportion is wide ([48%, 98%]), reflecting n=11. However, the asymmetry is robust to reasonable alternative classifications: if all 8 final "debatable" borderline cases (Appendix A) were reclassified as errors, the SUPPORT proportion remains 15/19 (79%) with two-sided p=0.019; under the expected 4/8 reclassification scenario, p=0.035. Nahum et al. (2025) observed a similar SUPPORT-skew pattern across factual-consistency benchmarks, lending external plausibility to the direction. We therefore describe this as a directional asymmetry warranting follow-up audit on a larger sample, rather than a confirmed systematic bias. With leaderboard gaps of 0.7–2.7 F1 between top systems, the binary correction-induced shifts of 1.7–3.8 F1 we observe across GPT-5.4 and Haiku 4.5 introduce noise of the same order as inter-system margins; the larger 3-way Haiku gain (+9.24) reflects 6 corrections that the binary schema cannot represent and is best read as an NEI-sensitive analysis rather than a leaderboard-comparable number. Whether the original-label rankings would actually reorder under corrected labels depends on whether top fine-tuned systems succeed and fail on the same items as our zero-shot models, which we do not establish here.

Six years without verification. Despite 600+ citations, our review of the dataset's GitHub repository (27 issues, none on annotation quality), Hugging Face page, and citing literature found zero prior label corrections or errata. This contrasts with other benchmarks: CoNLL-03 was re-annotated after two decades (Rucker and Akbik, 2023), and Vendrow et al. (2025) corrected 15 benchmarks simultaneously. We hypothesize that the "expert-annotated" label and biomedical domain specificity may have discouraged the kind of systematic verification that crowdsourced benchmarks routinely receive.

6 Conclusion

SciFact remains a useful benchmark, but its audited train and dev portions contain nontrivial annotation errors that can distort optimization targets and reported system comparisons. We document a directional asymmetry (9/11 dev errors mislabel SUPPORT; one-sided p=0.033, n=11) and show that correcting the dev labels raises binary zero-shot macro-F1 by 1.7–3.8 points across GPT-5.4 (mini, nano) and Claude Haiku 4.5—comparable to inter-system margins on the leaderboard. The 3-way Haiku gain is larger (+9.24) when the 6 mislabeled-as-NEI pairs are admitted; we report it as an additional NEI-sensitive analysis rather than as a direct leaderboard-comparable number. A claim-only contamination probe with Haiku 4.5 does not support label memorization as the main explanation for these gains. Rather than positioning our corrections as a definitive replacement, we recommend that the community treat them as a starting point and pursue an independent multi-annotator re-audit. We release the audit corrections, a blind-annotator packet for that review, and our reproduction scripts as a starting point.[6]

Limitations

We audit the dev and training sets only; the hidden test set may differ in error rate and error-type distribution, and we cannot directly verify it. *Manual verification is by one human annotator*, who was assisted on the 57 mini-flagged disagreements by a chat-mode frontier-LLM second opinion (§3.2) but not by an independent second human rater—a real concern given the original Cohen's κ=0.75 on the same task. To address this, we release a fully blind re-annotation packet for the entire dev set (no original gold, no LLM signal in the file the annotator sees) so any subsequent reviewer can independently re-audit and report inter-annotator agreement. Following Pavlick and Kwiatkowski (2019), some identified "errors"—particularly cross-species generalizations, borderline p=0.07 readings, and overgeneralization-type

[6] Corrected annotations, blind packet, and audit scripts are available at https://github.com/Kefez/scifact-audit-bionlp2026.

cases—may reflect legitimate methodological disagreement among clinical experts rather than objective labeling failures; the entity-mismatch errors (IDs 1216, 1274, 847) rely on the principle that findings in one organism do not automatically transfer to another, which some annotators may consider acceptable. The 8 final debatable cases (10 originally identified at first-pass review, with 2 subsequently upgraded to confirmed errors and counted in 11/209) are enumerated in Appendix A with a sensitivity analysis; case-by-case justifications are in the supplementary materials. The headline error rate stays under 10% even if all 8 were reclassified as errors. The impact analysis uses zero-shot models only on a single non-GPT family (Claude Haiku 4.5); fine-tuned systems and additional open-weight families are not evaluated and cannot be ruled out as differing in their response to the corrections. The contamination probe is a single negative test (claim-only with Haiku 4.5); a more extensive probe across open-weight models with known training data would strengthen the conclusion. Finally, our scope is a single benchmark; whether the directional SUPPORT-skew generalizes to other scientific-claim-verification corpora is an open empirical question.

Acknowledgments

We thank the BioNLP 2026 reviewers for detailed feedback that materially improved the statistical framing, cross-model verification, and contamination analysis in this paper.

References

Saso Cemerski, Jayajit Das, Emanuele Giurisato, Mary A. Markiewicz, Paul M. Allen, Arup K. Chakraborty, and Andrey S. Shaw. 2008. The balance between T cell receptor signaling and degradation at the center of the immunological synapse is determined by antigen quality. *Immunity*, 29(3):414–422.

Daria Esterházy, Ina Stützer, Haiyan Wang, Markus P. Rechsteiner, Jeremy Beauchamp, Heinz Döbeli, Hans Hilpert, Hugues Matile, Michael Prummer, Alexander Schmidt, and 1 others. 2011. Bace2 is a β cell-enriched protease that regulates pancreatic β cell function and mass. *Cell Metabolism*, 14(3):365–377.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pages 107–112.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Miloš Košprdić, Adela Ljajić, Darija Medvecki, Bojana Bašaragin, and Nikola Milošević. 2024. Scientific claim verification with fine-tuned NLI models. In *Proceedings of IC3K/KMIS*.

Fan-Tao Meng, Yan-Li Wang, Juan Liu, Jing Zhao, Rui-Yuan Liu, and Jiang-Ning Zhou. 2012. ApoE genotypes are associated with age at natural menopause in Chinese females. *Age*, 34(3):653–660.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are LLMs better than reported? Detecting label errors and mitigating their effect on model performance. In *Proceedings of EMNLP*.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of EMNLP*.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of NeurIPS*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the ACL*, 7:677–694.

Susanna Rucker and Alan Akbik. 2023. CleanCoNLL: A nearly noise-free named entity recognition dataset. In *Proceedings of EMNLP*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of NeurIPS Datasets and Benchmarks*.

Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. Do large language model benchmarks test reliability? *arXiv preprint arXiv:2502.03461*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of EMNLP*, pages 7534–7550.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022a. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of NAACL*.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of EMNLP*.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of ACL*.

Zhiwei Zhang, Jihua Fang, and Barbara Plank. 2021. Abstract, rationale, stance: A joint model for scientific claim verification. In *Proceedings of EMNLP*.

A Debatable Cases and Sensitivity Analysis

We enumerate the **8** SciFact dev claims that survived all review passes as "debatable but defensible" rather than confirmed errors. At an earlier triage pass, 10 cases were initially classified as debatable in the manual-review spreadsheets (`my_verdict == "DEBATABLE"` in the two false-alarm review files); on a second-pass review, 2 of those (claims 847 and 1274) were upgraded to confirmed errors and are part of the 11/209 figure. The 8 final debatable claim IDs are 57, 163, 431, 540, 982, 1041, 1132, 1150. Original gold-label split among the 8: 6 SUPPORT, 2 CONTRADICT (claims 57 and 1041 are CONTRADICT).

Sensitivity analysis (8 final debatables). If we instead reclassified all 8 as confirmed errors, the dev error rate rises to 19/209 = 9.1% (95% CI 5.6–13.8%). Under the expected 4/8 reclassification (3 of 4 added are SUPPORT, matching the 75% gold-split rate), the rate is 15/209 = 7.2% (95% CI 4.1–11.6%). The directional asymmetry persists in every scenario, and *strengthens* as more debatables are admitted: 9/11 SUPPORT-mislabel (82%, two-sided p=0.065) at 0/8 reclassified; 12/15 (80%, p=0.035) at the expected 4/8 case; 15/19 (79%, p=0.019) at 8/8. Per-case justifications for all 8 (and the 2 upgrades) are in the supplementary materials.

B Reproduction Artifacts

We release: (1) corrected gold labels for all 11 dev and 13 training errors as JSON; (2) a blind re-annotation packet for the 209 dev pairs (no original gold, no LLM signal) for inter-annotator agreement studies; (3) a 30-pair random sample of LLM-rejected pairs for false-negative-rate audit of the screen; (4) the Haiku 4.5 cross-model and contamination-probe scripts; (5) the statistical-analysis script that reproduces every CI and binomial test in this paper.

BioRAG: A Systematic Ablation Study of Retrieval Strategies for Biomedical Question Answering

Krushil Bhojani[1] [*] , **Mayank Waghmare**[1] , **Hima Bindu Nandyala**[1]
Contributors: Krupali Hirpara[2] , Shekhar Yadav[3] , Aakash Malhan[4]

[1]State University Of New York Polytechnic Institute , USA
[2]P. P. Savani University , INDIA
[3]University of North Texas , USA
[4]Arizona State University , USA
bhojank@sunypoly.edu

Abstract

Retrieval-Augmented Generation (RAG) systems for biomedical question answering require careful retrieval strategy selection, yet no systematic comparison exists across modern retrieval techniques evaluated with LLM-based metrics on biomedical benchmarks. We present BioRAG, a systematic ablation study comparing seven retrieval strategies on BioASQ-13b using four RAGAs metrics: faithfulness, answer relevancy, context precision, and context recall. Our corpus comprises 1,954 PubMed neurology abstracts indexed with domain-adaptive BioMedBERT embeddings. We find that hybrid BM25 plus dense retrieval with Reciprocal Rank Fusion achieves faithfulness of 0.534 and context recall of 0.507, improvements of 50% and 85% respectively over naive dense retrieval, with non-overlapping bootstrap confidence intervals across three random seed re-samples. We further demonstrate that Hypothetical Document Embeddings improve faithfulness by 14% but reduce context precision by 52%, revealing a faithfulness-precision tradeoff with direct clinical implications. No single strategy dominates all four metrics, indicating that retrieval strategy selection must be driven by application-specific clinical requirements. The complete pipeline operates without proprietary API fees using open-source models, ensuring full reproducibility.

1 Introduction

Biomedical question answering (QA) demands factual precision and reliable grounding in peer-reviewed literature, properties that are directly tied to patient safety in clinical deployments. Retrieval-Augmented Generation (RAG) addresses these requirements by conditioning answer generation on passages retrieved from a curated corpus, binding model outputs to verifiable evidence (Lewis et al., 2020). Despite rapid adoption in clinical applications, a fundamental question remains unanswered: *which retrieval strategy should a system designer select for a given clinical use case?*

Existing studies either propose a single retrieval method without ablating alternatives (Xiong et al., 2024), or evaluate on general-domain benchmarks that do not reflect the lexical complexity of biomedical text. Studies that do compare strategies on BioASQ (Nentidis et al., 2025) rely on lexical metrics such as F1 and ROUGE that do not measure grounding verification against retrieved context, which is the primary safety concern in clinical deployment. Existing pipelines depend on proprietary judge models, limiting reproducibility for researchers without institutional compute access.

We address this gap with BioRAG, a systematic ablation study comparing seven retrieval strategies on BioASQ-13b using RAGAs evaluation metrics (Es et al., 2024). The seven strategies range from naive dense retrieval and Hypothetical Document Embeddings (Gao et al., 2023) to cross-encoder reranking (Nogueira and Cho, 2019), multi-query decomposition (Ma et al., 2023), Self-RAG (Asai et al., 2023), and hybrid BM25 plus dense retrieval (Cormack et al., 2009); full descriptions appear in Section 3. Our key findings are: (1) hybrid BM25 plus dense retrieval achieves the highest faithfulness (0.53) and context recall (0.51), improving over the naive baseline by 50% and 85% respectively, with non-overlapping bootstrap confidence intervals confirming statistical significance; (2) Hypothetical Document Embeddings (Gao et al., 2023) improve faithfulness but reduce context precision by 52%, revealing a previously undocumented clinical tradeoff; and (3) no single strategy dominates all metrics, motivating evidence-based strategy selection. The complete pipeline operates without proprietary API fees using open-source models, ensuring full reproducibility of all findings.

[*]Krushil Bhojani is the corresponding author.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 104–114
July 3-4, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Biomedical RAG evaluation. Xiong et al. (2024) introduced MIRAGE, evaluating 41 corpus-retriever combinations on five biomedical QA datasets, and MedRAG, a retrieval framework supporting multiple corpora and snippet sizes for clinical QA. While comprehensive in scope, both systems measure accuracy on multiple-choice questions rather than open-ended faithfulness or contextual relevance. Thakur et al. (2021) introduced BEIR, a heterogeneous benchmark evaluating zero-shot transfer of retrieval models across multiple domains; hybrid retrieval emerged as a consistently strong baseline across the BEIR domains, consistent with our V7 findings in the biomedical setting. Tsatsaronis et al. (2015) introduced the BioASQ challenge, providing expert-curated biomedical QA pairs with gold-standard supporting snippets that enable context recall evaluation against human-verified references. Nentidis et al. (2025) describe the latest BioASQ challenge series, where participating systems typically combine BM25 retrieval with neural reranking evaluated via exact-match F1. Neither framework applies LLM-based evaluation metrics that directly measure answer grounding. Ozaki et al. (2025) examine confidence calibration in medical RAG across multiple model configurations, finding that certain models can discriminate whether retrieved documents support the correct answer. Our work complements these studies by providing the first systematic ablation of retrieval strategies using RAGAs metrics on BioASQ.

Retrieval strategy advances. Gao et al. (2023) proposed Hypothetical Document Embeddings (HyDE), generating a hypothetical answer passage and using its embedding as the retrieval query to bridge the query-document semantic gap. Nogueira and Cho (2019) demonstrated that cross-encoder reranking substantially improves retrieval precision by scoring query-document pairs jointly rather than independently, establishing the two-stage retrieval paradigm used in production search systems. Ma et al. (2023) showed that decomposing complex queries into focused sub-questions improves recall on multi-faceted questions. Asai et al. (2023) introduced Self-RAG, an iterative approach in which the model critiques its own answer confidence and re-retrieves if confidence is insufficient. Cormack et al. (2009) established Reciprocal Rank Fusion as a parameter-free method for combining rankings from multiple retrieval sources. Dense retrieval using dual encoders was established by Karpukhin et al. (2020), whose DPR model forms the conceptual basis for the bi-encoder component in V1–V5. BM25 scoring follows Robertson et al. (1995), which established the Okapi parameters k_1 and b used in our V7 Hybrid implementation. Our work evaluates the remaining strategies under a single biomedical evaluation framework for the first time.

LLM-based RAG evaluation. Es et al. (2024) proposed RAGAs, an automated evaluation framework measuring faithfulness, answer relevancy, context precision, and context recall using an LLM judge. Unlike lexical metrics, RAGAs directly assesses answer grounding and retrieval quality, dimensions essential for clinical decision support. We adopt RAGAs as our primary evaluation framework with a locally deployed `gemma3:12b` model as judge, ensuring reproducibility without proprietary API fees and without rate-limiting constraints that affect multi-variant evaluation.

3 Methods

3.1 Corpus and Indexing

We collected 1,954 PubMed abstracts on neurological diseases published between 2015 and 2024 via the NCBI Entrez API, using MeSH terms for neurology and neurological diseases. Abstracts were segmented into structured sections (Background, Methods, Results, Conclusions) using header detection, then chunked to 512 tokens with 64-token overlap. The 512-token chunk size matches the maximum input length of BioMedBERT, and the 64-token overlap preserves sentence context across chunk boundaries, a standard setting in biomedical IR. This yielded 10,425 indexed chunks stored in a persistent ChromaDB vector database.

All chunks are embedded using `microsoft/BiomedNLP-BiomedBERT-base-uncased` (Gu et al., 2021), a BERT variant pre-trained on 29 million PubMed abstracts. Domain-adaptive embeddings capture biomedical terminology that general-purpose models encode poorly. We chose BioMedBERT over general models such as `all-MiniLM-L6-v2` because biomedical terms such as "tau phosphorylation" and "amyloid-beta cascade" occupy semantically coherent regions in the BioMedBERT embedding space, improving retrieval quality on clinical text.

3.2 Evaluation Benchmark

We evaluate on BioASQ Training 13b (Nentidis et al., 2025), containing 5,389 expert-curated biomedical questions with gold-standard answers and supporting PubMed snippets. BioASQ provides gold context snippets per question, enabling computation of context recall against a human-verified reference. We filter questions to neurology-relevant topics using 25 domain keywords to align the benchmark with our corpus domain; the full keyword list is provided in Appendix D. We sample 100 questions per variant, selecting factoid and summary question types that require open-ended answer generation rather than binary yes/no responses.

3.3 Retrieval Variants

Figure 1 illustrates the complete BioRAG pipeline. We evaluate seven retrieval strategies of increasing complexity:

V1 – Naive Dense Retrieval (Lewis et al., 2020): Single-stage cosine similarity search, $\text{sim}(q, d) = (\mathbf{q} \cdot \mathbf{d})/(\|\mathbf{q}\|\|\mathbf{d}\|)$, in BioMedBERT embedding space. Returns top-k chunks directly. Serves as the performance baseline against which all other variants are evaluated.

V2 – HyDE (Gao et al., 2023): The LLM generates a hypothetical PubMed-style abstract that would answer the query. This hypothesis is embedded and used as the retrieval vector rather than the original question embedding, bridging the semantic gap between question phrasing and scientific document language.

V3 – Cross-Encoder Reranker (Nogueira and Cho, 2019): Two-stage retrieval. Stage one uses dense bi-encoder search to fetch $k \times 4$ candidates. Stage two applies `ms-marco-MiniLM-L-12-v2` cross-encoder reranking, which attends jointly to the query and each candidate passage to produce a relevance score.

V4 – HyDE + Reranker: Combines V2 and V3. Hypothesis-guided retrieval expands the semantic coverage of the candidate pool, and cross-encoder reranking then orders candidates by true relevance. Tests whether the two improvements are additive.

V5 – Multi-Query Decomposition (Ma et al., 2023): The LLM decomposes each question into three focused sub-queries. Candidates are retrieved independently for each sub-query, deduplicated by PMID, and reranked against the original question using the cross-encoder. Designed to improve recall on complex biomedical questions with multiple intents.

V6 – Self-RAG (Asai et al., 2023): Iterative retrieval with LLM confidence critique. The system retrieves, generates a candidate answer, and prompts the LLM to assess confidence as HIGH, MEDIUM, or LOW. If confidence is below HIGH, a follow-up query is generated and retrieval repeats for up to three rounds. Targets hallucination reduction through iterative grounding verification.

V7 – Hybrid BM25 + Dense (Cormack et al., 2009): Parallel retrieval via BM25Okapi sparse search and BioMedBERT dense search. Rankings are merged using Reciprocal Rank Fusion with constant $k = 60$:

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \quad (1)$$

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + r(d)} \quad (2)$$

where R is the set of rankers and $r(d)$ is the rank of document d in ranker r. The fused list is then reranked by the cross-encoder. BM25 captures exact biomedical term matches such as gene names and drug identifiers that dense retrieval misses when query and document phrasing differ.

Late-interaction models such as ColBERT (Khattab and Zaharia, 2020) are excluded because their token-level index structure requires a dedicated retrieval engine incompatible with the shared ChromaDB vector store used across all variants, which would confound retrieval strategy with retrieval infrastructure.

Table 1 summarises the computational complexity of each variant in terms of LLM calls per question, candidate pool size, and retrieval stages.

Variant	LLM calls	Pool size	Stages	Rerank
V1 Naive	0	k	1	No
V2 HyDE	1	k	2	No
V3 Reranker	0	$4k$	2	Yes
V4 HyDE+R	1	$4k$	3	Yes
V5 MultiQ	1	$12k$	3	Yes
V6 Self-RAG	1–3	$3k$	2–6	No
V7 Hybrid	0	$8k$	3	Yes

Table 1: Retrieval variant complexity. LLM calls excludes answer generation. Pool size uses $k = 5$ top results. V6 Self-RAG calls vary by confidence level reached.

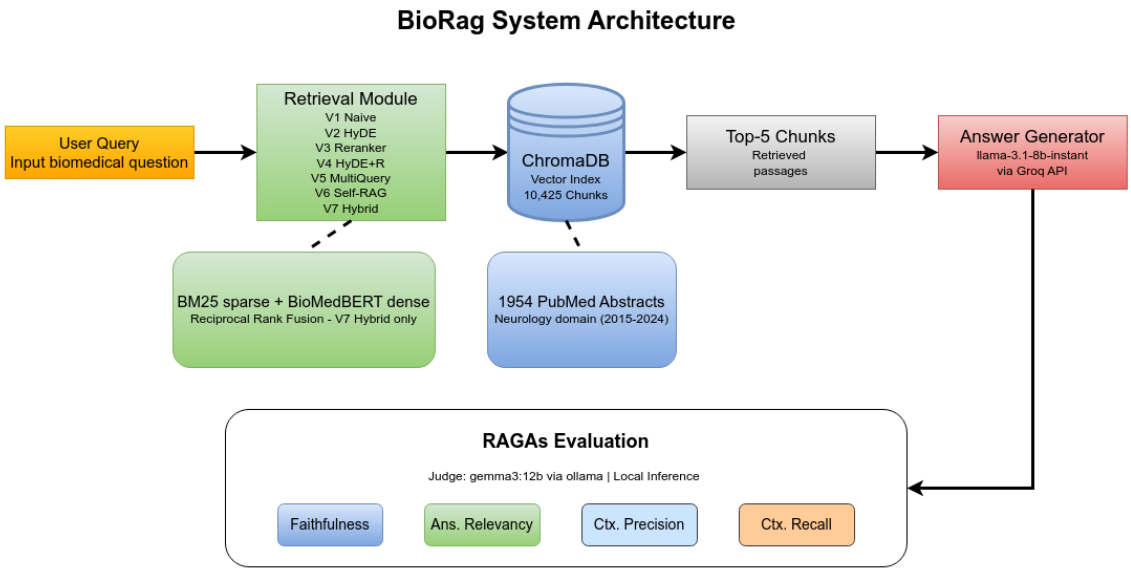


Figure 1: **BioRAG system architecture.** A biomedical query passes through one of seven retrieval variants (V1–V7), which retrieve the top-5 relevant chunks from a ChromaDB index of 10,425 BioMedBERT-embedded PubMed neurology abstracts. The answer generator (`llama-3.1-8b-instant` via Groq) produces a response from the retrieved context. The complete pipeline is evaluated using four RAGAs metrics with a locally deployed `gemma3:12b` judge.

3.4 Answer Generation

All answer generation uses `llama-3.1-8b-instant` via the Groq inference API. Answers are generated from the top-5 retrieved chunks using a prompt that instructs the model to synthesise from available context, falling back to general biomedical knowledge when context is incomplete. We use `temperature=0.1` and `max_tokens=512` for all variants to ensure comparable generation conditions. Total wall-clock time per question including answer generation ranges from approximately 5 seconds (V1 Naive) to 55 seconds (V6 Self-RAG) depending on Groq API latency.

3.5 Evaluation Setup

RAGAs evaluation (Es et al., 2024) uses `gemma3:12b` running locally via Ollama as the judge LLM, ensuring unlimited evaluation calls without API fees. We report four metrics per variant:

- **Faithfulness**: fraction of answer claims supported by retrieved context, measured by LLM-based claim extraction and verification.
- **Answer Relevancy**: degree to which the answer addresses the original question, measured via embedding similarity between the answer and synthetic questions generated from it.
- **Context Precision**: fraction of retrieved chunks that are relevant to the question, scored by the judge LLM per chunk.
- **Context Recall**: fraction of gold-standard information covered by the retrieved context, measured against BioASQ reference snippets.

To assess stability, we run each variant under three random seeds (42, 123, 456) and compute 95% bootstrap confidence intervals using 1,000 resamples. We additionally run a hyperparameter sensitivity analysis varying $k \in \{3, 5, 10\}$ for V1, V3, and V7. V6 Self-RAG is excluded from both analyses because its results are confounded by Groq free-tier rate limiting rather than algorithmic behaviour; including it would produce artificially stable low scores reflecting infrastructure constraints rather than retrieval quality. All other five variants plus V7 are included in the significance analysis. Reported scores in Table 2 correspond to the primary run with $k = 5$ and seed 42; the k-sensitivity results in Table 6 use seed 123 and reflect question sampling variance across seeds, as reported in Table 3.

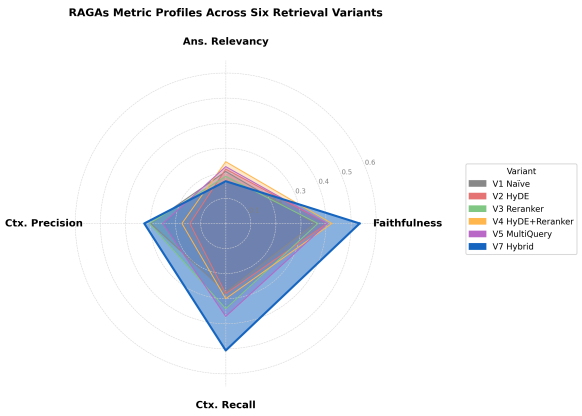


Figure 2: RAGAs metric profiles across six retrieval variants (V6 excluded from this chart due to infrastructure confounding; see Section 6). V7 Hybrid (dark blue) occupies the largest overall area, winning faithfulness, context precision, and context recall. V4 HyDE plus Reranker leads answer relevancy. No single variant dominates all four axes.

3.6 Implementation Details

All experiments run on a single consumer GPU with 8GB VRAM. BioMedBERT embeddings are computed locally via HuggingFace Transformers. The cross-encoder reranker uses a candidate pool of $k \times 4 = 20$ passages. BM25 uses Okapi parameters $k_1 = 1.5$ and $b = 0.75$. Groq free-tier inference provides 30 requests per minute and 500,000 tokens per day. A cooldown of 4 seconds between API calls prevents rate-limit exhaustion during evaluation. All code, evaluation scripts, and results are publicly available to support reproducible biomedical RAG research.

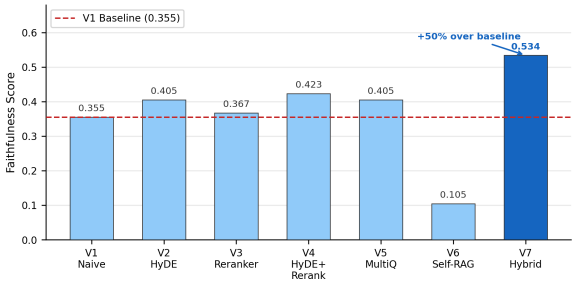


Figure 3: Faithfulness scores across all variants including V6 Self-RAG (seed 42, $k = 5$). V7 Hybrid achieves 0.534, a 50% improvement over the V1 Naive baseline (dashed line). V6 Self-RAG collapses to 0.105 due to Groq free-tier rate-limiting constraints; full V6 results appear in Appendix E.

4 Results

Table 2 presents RAGAs scores for six retrieval variants evaluated on 100 BioASQ-13b neurology questions (seed 42, $k = 5$); V6 Self-RAG is reported separately in Appendix E. Figure 3 shows faithfulness scores across all variants. Figure 2 presents the complete metric profile for V1–V5 and V7 across all four RAGAs dimensions. For reference, Es et al. (2024) report faithfulness scores of 0.4–0.6 on general-domain QA using GPT-4 as judge with a stronger model; our scores fall within this range despite using a smaller local judge on a domain-specific corpus, indicating reasonable calibration.

Hybrid retrieval dominates. V7 Hybrid achieves the highest faithfulness (0.534), context precision (0.325), and context recall (0.507), winning three of four metrics. The combination of BM25 sparse retrieval with BioMedBERT dense retrieval via Reciprocal Rank Fusion captures complementary signals: BM25 matches exact biomedical terms such as gene names and drug identifiers, while dense retrieval handles semantic paraphrasing. The 50% faithfulness improvement over V1 (0.355 to 0.534) and 85% context recall improvement (0.273 to 0.507) establish V7 as the strongest variant for biomedical RAG. These gains are consistent with findings from production search systems where hybrid retrieval outperforms single-signal approaches (Cormack et al., 2009).

HyDE introduces a faithfulness-precision tradeoff. V2 HyDE improves faithfulness over V1 (0.405 vs 0.355, +14%) but reduces context precision substantially (0.143 vs 0.300, −52%). Hypothesis generation shifts retrieval toward the answer embedding space, improving grounding but expanding the candidate pool with plausible yet

Variant	Faithfulness↑	Ans. Rel.↑	Ctx. Prec.↑	Ctx. Recall↑	Latency
V1 Naive Dense	0.355	0.207	0.300	0.273	46ms
V2 HyDE	0.405	0.217	0.143	0.280	1370ms
V3 Reranker	0.367	0.188	0.304	0.340	388ms
V4 HyDE+Reranker	0.423	**0.247**	0.175	0.300	1795ms
V5 MultiQuery	0.405	0.227	0.251	0.371	1611ms
V7 Hybrid	**0.534**	0.169	**0.325**	**0.507**	1551ms

Table 2: RAGAs evaluation results on 100 BioASQ-13b neurology questions using `gemma3:12b` as judge (seed 42, $k = 5$). Bold indicates best per metric. Latency reflects mean retrieval time only, excluding answer generation. V6 Self-RAG is excluded from this table due to infrastructure confounding; see Appendix E for full V6 results.

Variant	Metric	Mean	Std	95% CI lower	95% CI upper
V7 Hybrid	Faithfulness	**0.585**	0.021	0.565	0.607
V2 HyDE	Faithfulness	0.463	0.052	0.406	0.506
V4 HyDE+Reranker	Faithfulness	0.448	0.060	0.378	0.486
V5 MultiQuery	Faithfulness	0.412	0.036	0.372	0.442
V1 Naive	Faithfulness	0.365	0.033	0.329	0.393
V3 Reranker	Faithfulness	0.352	0.022	0.327	0.366
V7 Hybrid	Ctx. Recall	**0.478**	0.042	0.430	0.503
V1 Naive	Ctx. Recall	0.372	0.018	0.353	0.388
V3 Reranker	Ctx. Recall	0.340	0.015	0.326	0.356

Table 3: Bootstrap confidence intervals (95%, 1,000 resamples) across three random seeds (42, 123, 456) for faithfulness and context recall. Means differ from Table 2 single-seed values due to question sampling variance across seeds (standard deviations confirm this). V7 Hybrid confidence intervals do not overlap with any other variant on either metric, confirming statistical significance of the reported gains. V2 and V5 faithfulness intervals overlap, indicating no reliable difference between those two variants. V6 Self-RAG excluded due to infrastructure confounding.

imprecise passages. V4 HyDE plus Reranker partially recovers precision (0.175) while achieving the highest answer relevancy (0.247), suggesting that reranking mitigates but does not eliminate the HyDE precision penalty. This tradeoff has not been previously documented on BioASQ and represents a key finding for clinical system designers who must balance faithfulness against retrieval precision.

Reranking improves context recall. V3 Reranker achieves the highest context recall among single-enhancement variants (0.340 vs 0.273 baseline, +24.5%). Cross-encoder joint scoring retrieves passages that are genuinely relevant to the question rather than merely semantically similar in embedding space. V5 MultiQuery extends this further by decomposing questions into sub-queries, achieving 0.371 context recall, the second highest across all variants, by covering multiple facets of complex biomedical questions.

No single strategy dominates all metrics. V7 leads faithfulness, context precision, and context recall while V4 leads answer relevancy. This multi-winner result has direct clinical implications: safety-critical applications requiring grounded answers should prefer V7 Hybrid, while end-user-facing systems prioritising direct question relevance should prefer V4 HyDE plus Reranker. Latency-constrained deployments should prefer V1 Naive (46ms) or V3 Reranker (388ms).

Self-RAG is not viable on rate-limited APIs. V6 Self-RAG, reported separately in Appendix E due to infrastructure confounding, achieves faithfulness of 0.105 and latency of 48,730ms under Groq free-tier constraints. The iterative critique loop generates multiple sequential API calls per question, exhausting Groq free-tier rate limits (30 requests per minute) and introducing 10–30 second wait periods between retrieval rounds. This stalling amplifies retrieval noise rather than resolving uncertainty, as the model accumulates low-confidence context across rounds. Self-RAG requires dedicated inference capacity or local LLM deployment to function as intended.

Statistical significance. Table 3 reports bootstrap confidence intervals across three random seeds. V7 Hybrid faithfulness mean of 0.585 with 95% CI [0.565, 0.607] does not overlap with any other variant, confirming that V7 gains are statistically reliable. V2 HyDE and V5 MultiQuery faithfulness intervals overlap, indicating no reliable difference between those two strategies. Means in Table 3 differ from Table 2 single-seed values because each seed samples a different question subset; the standard deviations quantify this sampling variance.

5 Discussion

5.1 Clinical Deployment Implications

Our results provide concrete guidance for practitioners designing biomedical RAG systems. The choice of retrieval strategy should be driven by the primary clinical objective rather than a single performance number. For safety-critical applications such as clinical decision support, where hallucinated claims can directly harm patients, V7 Hybrid maximises faithfulness (0.534) and should be the default choice. For patient-facing QA systems where answer relevance and directness matter most, V4 HyDE plus Reranker achieves the highest answer relevancy (0.247). For evidence review applications such as systematic literature search, V7 also dominates with context recall of 0.507, retrieving over half the gold-standard evidence from BioASQ reference snippets.

Table 4 maps each clinical use case to the recommended variant and primary metric.

Clinical Use Case	Recommended	Key Metric
Decision support	V7 Hybrid	Faithfulness
Patient-facing QA	V4 HyDE+R	Ans. Rel.
Systematic review	V7 Hybrid	Ctx. Recall
Real-time interface	V3 Reranker	Latency
High-throughput	V1 Naive	Latency

Table 4: Hypothesized retrieval strategy mappings by clinical use case, based on RAGAs metric performance. Clinical validation is required before deployment.

5.2 The HyDE Faithfulness-Precision Tradeoff

The faithfulness-precision tradeoff introduced by HyDE carries a specific clinical implication. Hypothesis generation improves answer grounding by aligning retrieval with the answer embedding space, but it simultaneously retrieves a broader, noisier candidate pool. In clinical settings where precision is critical, for example retrieving the correct drug dosage rather than broadly related pharmacology, HyDE alone is counterproductive. V4 HyDE plus Reranker partially addresses this by filtering the expanded pool, making it a better choice when both grounding and precision matter.

5.3 Latency and Production Tradeoffs

Latency varies across variants from 46ms for V1 Naive to 48,730ms for V6 Self-RAG. These figures reflect retrieval time only; answer generation via Groq adds approximately 3–8 seconds per question across all variants. For real-time clinical applications, V3 Reranker offers the best accuracy-latency tradeoff: it achieves the highest context recall among non-hybrid variants (0.340) at only 388ms retrieval time, making it suitable for interactive QA interfaces. V7 Hybrid delivers the strongest overall performance at 1,551ms retrieval time, acceptable for asynchronous clinical workflows but potentially prohibitive for real-time decision support. V1 Naive at 46ms remains appropriate for high-throughput screening applications where speed outweighs accuracy. Groq answer generation adds approximately 3–8 seconds to all variants, meaning end-to-end latency differences between V1 and V3 are less differentiating in practice than retrieval times alone suggest.

5.4 Reproducibility

BioRAG operates without proprietary API fees throughout the evaluation pipeline: BioMedBERT embeddings run locally, ChromaDB requires no cloud infrastructure, Groq provides free-tier answer generation, and Ollama hosts the RAGAs judge locally. Researchers at institutions without GPU clusters or API budgets can replicate all experiments using consumer hardware. The entire evaluation pipeline for all seven variants runs in approximately 15 hours on a single 8GB GPU.

6 Limitations

Our corpus of 1,954 neurology abstracts covers a single biomedical subdomain, while BioASQ spans all of biomedicine. A domain mismatch diagnostic on the 100 evaluation questions found that 2 of 100 questions (2.0%) returned no relevant corpus material, confirmed by V1 Naive aggregate faithfulness (0.355) remaining above the mismatch threshold. This low mismatch rate indicates that reduced absolute RAGAs scores reflect retrieval qual-

ity differences rather than corpus coverage failures, supporting the validity of relative comparisons.

RAGAs evaluation with `gemma3:12b` as judge has not been validated against human expert assessments on biomedical QA. This is a primary methodological concern: an unvalidated judge introduces unknown calibration bias, and ranking differences between close variants such as V2 and V5 may not reflect true retrieval quality differences. Future work should conduct human annotation on a subset of questions to validate judge rankings against expert assessments. Evaluation uses a smaller judge than GPT-4-class models, which may introduce calibration differences relative to stronger judges; however, this is necessary for reproducibility without proprietary API costs.

The answer generation prompt permits fallback to general biomedical knowledge when context is incomplete, meaning faithfulness scores partially reflect the generator's parametric knowledge rather than retrieval quality alone. Future work should compare strict context-only prompts to isolate retrieval contribution. Answer generation uses `llama-3.1-8b-instant`, a smaller model than those used in comparable studies, which may underestimate the performance ceiling of each retrieval strategy when paired with a stronger generator. Generator and retriever quality interact in ways this study cannot fully disentangle with a single generator model. V6 Self-RAG performance is specifically constrained by free-tier API rate limits rather than algorithmic limitations; with dedicated inference, Self-RAG would operate at comparable latency to other LLM-dependent variants. The evaluation does not include an error analysis categorising failure modes by question type; future work should examine whether retrieval failures concentrate in specific question categories such as multi-hop or causal questions. Future work should validate findings with a larger multi-domain corpus, human judge validation, a stronger generator, and unconstrained inference for Self-RAG.

7 Conclusion

We presented BioRAG, a systematic ablation study comparing seven retrieval strategies for biomedical question answering on BioASQ-13b. Hybrid BM25 plus dense retrieval with Reciprocal Rank Fusion achieves the strongest overall performance, improving faithfulness by 50% and context recall by 85% over naive dense retrieval, with bootstrap confidence intervals confirming these gains are statistically reliable. HyDE consistently improves faithfulness but introduces a 52% context precision penalty, a tradeoff with direct implications for clinical RAG system design that has not been previously documented on BioASQ. No single strategy dominates all four RAGAs metrics, demonstrating that retrieval strategy selection must be driven by application-specific clinical requirements. The complete evaluation pipeline is released to support reproducible biomedical RAG research in line with BioNLP 2026's emphasis on evaluation frameworks and reproducibility of findings.

Alongside the evaluation pipeline, we release an interactive Streamlit demo enabling practitioners to compare all seven retrieval variants on custom queries, supporting the translation of our findings into biomedical NLP practice.

Acknowledgements

The authors thank Shekhar Yadav (University of North Texas), Aakash Malhan (Arizona State University). for their contributions to the statistical experimental design and significance testing methodology. This research was conducted independently and received no external funding.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Advances in Neural Information Processing Systems*, volume 36.

Gordon V Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference*, pages 758–759.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 150–158.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1762–1777.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng

Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 1–5.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, and Georgios Paliouras. 2025. Overview of BioASQ tasks 13b and Synergy13 in CLEF2025. In *Conference and Labs of the Evaluation Forum*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5723–5729.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2025. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 1–17.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. *NIST Special Publication*, 500:109–126.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Muthukrishnan, Martin Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, and Dimitris Polychronopoulos. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

A Prompt Templates

All prompts use `llama-3.1-8b-instant` via Groq API with `temperature=0.1`.

HyDE Hypothesis Generation (V2, V4).

> *Write a PubMed-style abstract that directly answers the following biomedical question. Be specific and use scientific terminology. Question: {question}*

Multi-Query Decomposition (V5).

> *Generate three different search queries to retrieve relevant biomedical literature for answering this question. Each query should focus on a different aspect. Question: {question}. Output format: Query 1: ... Query 2: ... Query 3: ...*

Self-RAG Confidence Critique (V6).

> *Given the question and retrieved context, assess your confidence in the answer. Respond in JSON: {"confidence": "HIGH/MEDIUM/LOW", "missing": "...", "followup_query": "..."}. Question: {question}. Context: {context}. Candidate answer: {answer}*

Answer Generation (all variants).

> *You are a biomedical expert. Answer the question using the provided context. Prioritise information from the context. If the context is relevant but incomplete, synthesise an answer from what is available. If the context is entirely unrelated, draw on general biomedical knowledge. Question: {question}. Context: {context}*

B Example Evaluation Cases

Table 5 presents three representative BioASQ questions from our evaluation set with answers generated by V1 Naive and V7 Hybrid, illustrating the practical impact of retrieval strategy on answer quality.

Question	V1 Naive answer	V7 Hybrid answer
What is the association between neuroticism and Alzheimer's disease risk?	High levels of neuroticism have been associated with increased risk of cognitive decline in older adults, though the mechanisms remain unclear.	High neuroticism is associated with significantly increased Alzheimer's disease risk and more advanced neuropathology, including greater amyloid burden and tau phosphorylation, independent of depression.
What is the association between adiponectin and migraine?	Adiponectin is an adipokine involved in metabolic regulation and has been studied in various neurological conditions.	Adiponectin levels are elevated in episodic migraine patients compared to controls, suggesting a role in migraine pathophysiology potentially linked to neurogenic inflammation.
Does H. pylori infection increase risk for ischemic stroke?	H. pylori is a gastrointestinal pathogen that has been associated with systemic inflammation in some studies.	Evidence is conflicting, but several studies report that H. pylori infection is associated with increased ischemic stroke risk, potentially mediated through inflammatory and prothrombotic mechanisms.

Table 5: Representative BioASQ evaluation examples comparing V1 Naive and V7 Hybrid answers. V7 consistently produces more specific, evidence-grounded responses due to BM25 capturing exact biomedical terminology.

C Hyperparameter Sensitivity

Table 6 reports faithfulness and context recall for V1 Naive, V3 Reranker, and V7 Hybrid across $k \in \{3, 5, 10\}$ retrieved chunks. These runs use seed 123 and a freshly sampled question subset; values differ from Table 2 (seed 42) due to question sampling variance, which is quantified in Table 3. V7 Hybrid maintains the highest faithfulness at $k = 5$ and $k = 10$, with context recall improving from 0.368 to 0.658 as k increases. V3 Reranker shows non-monotonic behaviour, peaking at $k = 10$ for faithfulness (0.463) but dropping context recall at $k = 10$ (0.238). V1 Naive faithfulness increases steadily with k, consistent with the finding that more context benefits simpler retrieval more than strategies that already filter aggressively. These results confirm that $k = 5$ is a conservative setting and that relative rankings between variants are preserved across the tested range.

D Neurology Domain Keywords

The following 25 keywords were used to filter BioASQ Training 13b questions to the neurology domain: *alzheimer, tau, dopamine, stroke, epilepsy, parkinson, dementia, neuron, cortex, hippocampus, amyloid, synapse, neurotransmitter, seizure, migraine, multiple sclerosis, neuropathy, glioma, cerebral, meningitis, encephalitis, neurodegeneration, myelin, axon, spinal*.

E V6 Self-RAG Infrastructure Results

Table 7 reports V6 Self-RAG results for completeness. These results are excluded from Table 2 because they reflect Groq free-tier rate-limiting constraints rather than algorithmic performance. The iterative critique loop generates multiple sequential API calls per question, exhausting the 30 requests-per-minute limit and introducing 10–30 second wait periods between rounds. With dedicated local inference, Self-RAG would operate at latency comparable to V4 and V5.

Variant	k	Faith.	Ctx. Prec.	Ctx. Rec.	Lat. (ms)
V1 Naive	3	0.323	0.276	0.267	48
V1 Naive	5	0.404	0.283	0.288	54
V1 Naive	10	0.440	0.282	0.332	52
V3 Reranker	3	0.381	0.331	0.298	215
V3 Reranker	5	0.359	0.310	0.352	319
V3 Reranker	10	0.463	0.299	0.238	781
V7 Hybrid	3	0.431	0.295	0.368	917
V7 Hybrid	5	**0.476**	**0.325**	0.507	1588
V7 Hybrid	10	0.474	0.294	**0.658**	3240

Table 6: Hyperparameter sensitivity across $k \in \{3, 5, 10\}$ using seed 123. Values differ from Table 2 (seed 42) due to question sampling variance across seeds; see Table 3 for cross-seed standard deviations. Bold indicates best per metric across all rows.

Variant	Faith.	Ans.Rel.	Ctx.Prec.	Ctx.Rec.	Latency
V6 Self-RAG	0.105	0.050	0.286	0.310	48730ms

Table 7: V6 Self-RAG results under Groq free-tier rate-limiting. Scores reflect infrastructure constraints, not algorithmic capability. Included for completeness only.

Post Hoc Agentic Refinement for Improving Precision in Multilingual Clinical Text De-identification

Justin Xu[1,2,*], **Alistair E W Johnson**[2], **Thomas Lin**[2],
David W Eyre[1], **Rodolfo Quispe**[2]
[1]University of Oxford, [2]Microsoft Health & Life Sciences

Abstract

De-identification systems prioritize recall to protect privacy, but excessive over-tagging reduces data utility. We propose an agentic refiner that reviews high-recall annotations using lightweight tools (validation functions, adaptive context retrieval, persistent to-do state, and modular review skills) to improve precision while minimizing recall loss. Experiments across three multilingual datasets show that the agent achieves significant improvements to binary precision. To support fine-grained analysis, we further introduce a synthetic error dataset of common and systemic failure modes, on which the agent corrects >99% of injected errors in the medical datasets. Our results suggest that agent-based refinement provides a flexible and effective mechanism for improving de-identification precision as a modular extension to existing high-recall systems.

1 Introduction

De-identification of text aims to detect and remove or mask Protected Health Information (PHI) in clinical data and, more broadly, Personally Identifiable Information (PII) in unstructured text. This task is a prerequisite for data sharing, secondary use, and deployment of downstream NLP systems in sensitive domains. From a privacy perspective, recall is paramount: failing to identify a true PHI/PII instance risks irreversible disclosure, whereas over-tagging benign text is less harmful in terms of privacy. Consequently, many de-identification systems are optimized to maximize recall, often at the expense of precision (Uzuner et al., 2007; Stubbs and Uzuner, 2015; Li et al., 2024). However, in many real-world settings, practitioners may be willing to trade small amounts of recall for meaningful improvements in precision, particularly when working within systems with additional safeguards for patient privacy. Excessive over-tagging degrades data utility, harms readability, and can negatively impact applications such as clinical coding, summarization, retrieval, and generative model training. The ideal de-identification system should thus maintain both high recall and high precision – a balance that remains challenging to achieve in scenarios requiring nuanced contextual reasoning (Yang et al., 2019; Li et al., 2024).

In this work, we study post hoc refinement of de-identification outputs as a complement to existing high-recall systems. Rather than introducing a new end-to-end model, we assume an initial set of PHI/PII annotations and focus on improving precision. As a baseline, we consider a simple LLM-as-a-judge approach that reviews annotation spans in context. We then introduce an agentic refiner, designed to reason over annotations using Model Context Protocol (MCP) (Hou et al., 2025) tools and iterative decision-making, enabling more consistent refinement. We additionally construct a synthetic error dataset for fine-grained evaluation of error detection and correction.

Our contributions can be summarized as follows:

1. We propose a modular agent-based refinement pipeline that can be attached to the output of any existing de-identification system to improve precision without retraining (Figure 1).
2. We evaluate this approach across three complementary datasets spanning Spanish clinical notes, German medical transcripts, and English general-domain text with PII (Table 1).
3. We introduce methods to create a synthetic error dataset for de-identification refinement, evaluating error detection (Section 5).

2 Related Works

Early de-identification systems relied on rule-based methods using handcrafted patterns and dictionaries (Neamatullah et al., 2008). Later work framed

[*]Work done during internship at Microsoft

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 115–127
July 3-4, 2026 ©2026 Association for Computational Linguistics

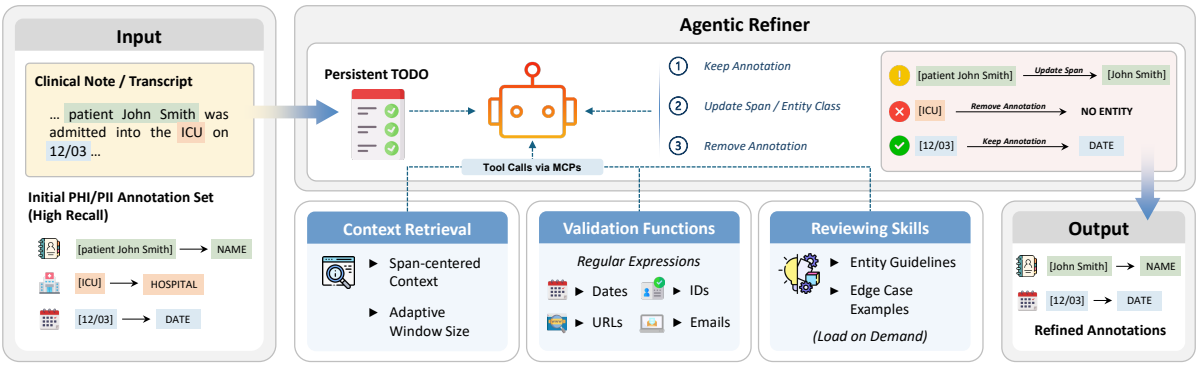


Figure 1: Schematic of Post Hoc Agentic Refinement of Annotations in Clinical Text De-identification.

de-identification as a supervised sequence labeling task, adopting Transformer-based models (BERT and its clinical variants (Alsentzer et al., 2019; Johnson et al., 2020; Huang et al., 2020; Lee et al., 2025; Sounack et al., 2025)). Prior efforts have also combined neural models with external knowledge to generate higher-precision silver data (Tedeschi et al., 2021). More recently, LLMs have been applied to de-identification in zero/few-shot settings, demonstrating acceptable recall and flexibility (Singh et al., 2025; Dai et al., 2025; Zhu et al., 2025; Ding et al., 2025; Aghakasiri et al., 2025). However, LLM-based methods often over-predict PHI/PII and exhibit inconsistent labeling, which is highly sensitive to linguistic nuances in the provided prompt. Agentic systems have also been explored for end-to-end information extraction and clinical reasoning (Yao et al., 2023; Shen et al., 2023; Lu et al., 2025; Lin et al., 2025), but remain costly and difficult to control for de-identification.

In contrast, our work focuses on agent-based refinement rather than replacement. By operating only on an existing set of annotations, our approach provides guardrails to an otherwise open-ended NER task. This modular design allows the agent to be integrated as a precision-enhancing layer on top of high-recall de-identification pipelines.

3 Experimental Design

3.1 Problem Definition

Let a document be represented as a token sequence $D = \{w_1, \ldots, w_n\}$. An upstream de-identification system produces an initial set of PHI/PII annotations $\mathcal{A} = \{a_1, \ldots, a_K\}$, where each annotation $a_i = (s_i, e_i, c_i, t_i)$ consists of a start index s_i, end index e_i, an entity class $c_i \in \mathcal{C}$, and a surface text span $t_i = \{w_{s_i}, \ldots, w_{e_i}\}$.

Given the document D and an associated annotation set $\mathcal{A}$, the refinement task aims to produce an updated set $\mathcal{A}^*$ by evaluating each a_i independently. For a given annotation, the refiner decides whether to (i) retain it unchanged, (ii) update its span boundaries (s_i, e_i), (iii) modify its class c_i, or (iv) remove it entirely if it does not correspond to PHI/PII. The refiner does not introduce new annotations; its sole role is to improve the quality of an existing high-recall annotation set.

This formulation treats de-identification as a contextualized span validation problem rather than end-to-end sequence labeling, enabling refinement to be applied as a post hoc layer on top of arbitrary upstream systems.

3.2 Datasets and Evaluation Metrics

We evaluate annotation refinement across three multilingual datasets. For all datasets, refinement is applied to an initial set of PHI/PII annotations produced by GPT-5 (`gpt-5-2025-08-07`) in a few-shot setting, with entity class-specific guidelines (prompt available in Appendix A.1).

CARMEN-ES (5,614 Total PHI). Derived from the Spanish subset of the *CARMEN-I* corpus of 2,000 documents (Farre Maduell et al., 2024), comprising of discharge summaries, referrals, and radiology reports from Hospital Clínic of Barcelona (March 2020 – March 2022). The collection primarily covers COVID-19 patients with diverse comorbidities. Documents in Catalan or mixed Spanish-Catalan were excluded, and high-quality gold labels were previously validated by experts.

Transcripts-DE (238 Total PHI). Conversational clinical encounters in the German language simulated by medical actors and grounded in real-world patient cases from a hospital in Germany.

WikiPII-EN (24,145 Total PII). Info-boxes with people and animal descriptions derived from English Wikipedia articles (Wang et al., 2018). This represents a general-domain de-identification setting and reflects common non-clinical applications.

Evaluation. Our primary focus is binary PHI/PII detection. An annotation is considered correct if its span overlaps with a gold PHI/PII span, regardless of entity class. We report precision, recall, and F1 where precision measures the proportion of predicted PHI/PII tokens that are correct and recall measures the proportion of gold PHI/PII tokens that are recovered.

3.3 LLM Refiner Baseline

As a baseline, we implement an LLM-based refiner using GPT-5 (`gpt-5-2025-08-07`, medium verbosity and reasoning effort). For each annotation a_i, the model is prompted with the document D, the surface text span t_i from (s_i, e_i), its class c_i, and general PHI/PII labeling guidelines, and is asked to determine whether the annotation should be kept, modified, or removed (prompt available in Appendix A.2). Few-shot examples are provided to encourage consistent behavior.

As seen in "Post-Refinement (GPT-5, Single Call)" of Table 1, while this approach improves binary precision, it removes many valid PHI/PII spans, leading to substantial recall degradation. In addition, the method requires one large API call per annotation with long contextual inputs, motivating a more optimized structured refinement approach.

4 Agentic Refiner

We model annotation refinement as an agentic decision process operating over a document D and its initial annotation set $\mathcal{A} = \{a_1, \ldots, a_K\}$. The agent similarly processes annotations sequentially and produces a refined set $\mathcal{A}^*$.

At each step, the agent maintains a state $\sigma_i = (D, a_i, h_i)$, where $a_i = (s_i, e_i, c_i, t_i)$ is the current annotation under review and h_i denotes the agent's internal memory, including prior decisions and intermediate observations. Given σ_i, the agent selects an action o_i from a discrete refinement space corresponding to retaining, modifying, or removing a_i. Actions may involve tool calls executed through MCP servers, after which the agent updates its state and proceeds to the next annotation.

Formally, the agent implements a policy $\pi(o_i \mid \sigma_i)$ that interleaves natural language reasoning with tool use, producing $\mathcal{A}^*$ after all annotations have been reviewed. Unlike the LLM baseline, the agent's decisions are informed by persistent memory and structured tool outputs, enabling more consistent refinement across long documents. Appendix B highlights representative agentic outputs.

Persistent To-Do List. The agent initializes a persistent to-do list containing all annotations in $\mathcal{A}$ for a given document and marks items as completed as refinement progresses. This externalized memory allows progress tracking over long clinical notes, which may contain several dozen annotations with overlapping surface forms. Similar task decomposition strategies have been shown to improve reliability and long-horizon performance (Yao et al., 2023; Shen et al., 2023).

Pattern-Based Validation Tools. We provide deterministic validation functions based on regular expressions for entity types with well-defined surface forms, such as dates, email addresses, URLs, and identifiers. When applicable, the agent can invoke these tools to validate or reject an annotation without extended reasoning. Offloading high-confidence decisions to symbolic components can reduce hallucinations and improve grounding in hybrid neuro-symbolic and tool-augmented systems (Gao et al., 2023; Schick et al., 2023).

Adaptive Context Retrieval. Rather than always conditioning on the full document D, the agent can dynamically retrieve a window of context around the annotation span, selecting a variable number of tokens to the left and right of t_i based on (s_i, e_i). This flexibility is particularly important for conversational transcripts, where relevant context may span multiple turns of varying length. Adaptive context selection reduces unnecessary token usage while preserving critical discourse information.

Modular Review Skills. Finally, the agent can load specialized review *skills* on demand (Zhang et al., 2025) – modular guideline documents associated with groups of related entity classes. These skills include entity definitions (identical to the guidelines provided to the baseline refiner) and examples of correct and incorrect annotations (including edge cases). The agent decides when to load a skill during its reasoning process, allowing targeted access to entity class-specific knowledge and enables incremental future development.

Dataset	Precision ↑	Recall ↑	F1 ↑
Original GPT-5 Annotations			
CARMEN-ES	90.76	96.75	93.65
Transcripts-DE	56.01	92.02	69.67
WikiPII-EN	83.03	96.85	89.41
Post-Refinement (GPT-5, Single Call)			
CARMEN-ES	94.31 (+3.55)	93.97 (-2.78)	94.14 (+0.49)
Transcripts-DE	69.96 (+13.95)	83.19 (-8.83)	75.97 (+6.30)
WikiPII-EN	90.72 (+7.69)	92.03 (-4.82)	91.37 (+1.96)
Post-Refinement (Agentic)			
CARMEN-ES	97.95 (+7.19)	95.39 (-1.36)	**96.65** (+3.00)
Transcripts-DE	83.68 (+27.67)	87.39 (-4.63)	**85.49** (+15.82)
WikiPII-EN	94.92 (+11.89)	88.79 (-8.06)	**91.75** (+2.34)

Table 1: Binary performance metrics (precision, recall, and F1) before and after refinement across datasets. Values in brackets denote changes relative to pre-refinement (original annotations) performance metrics.

Across all datasets, agentic refinement with GPT-5 (`gpt-5-2025-08-07`, medium verbosity and reasoning effort) yields substantially larger precision gains than the baseline, while better controlling recall degradation (leading to optimal F1). On CARMEN-ES, the agent more than doubles the baseline's precision gain, while incurring a smaller recall drop (over 6× F1 boost). For the German transcripts, agentic refinement also achieves more than 2× the baseline precision gain at a lower cost in recall points, reflecting a favorable precision-recall trade-off in a challenging conversational setting. On WikiPII, the agent again delivers stronger precision gains and overall F1, albeit with a larger recall reduction, indicating that aggressive refinement is more likely to remove borderline entities in general-domain text.

5 Error Dataset

To facilitate a precise and controlled evaluation of annotation refinement, we construct a synthetic error dataset in which the exact location and type of each error is known. This allows us to directly measure how effectively a refiner identifies and corrects annotation errors (Table 2). Starting from reference annotations, we inject realistic errors into medical datasets using three complementary strategies.

Regular Expressions. First, we define a set of regular expression-based rules that inject common and well-known de-identification errors. These rules simulate mistakes frequently observed in both statistical and neural de-identification systems. In total, 10 such rules were specified and applied deterministically; the full list is provided in Appendix C.

Systemic Errors via Clustering. Second, we identify systemic error patterns made by LLM-based de-identification systems. We collect incorrect annotations produced by LLMs along with their surrounding context and encode them using a sentence transformer (Reimers and Gurevych, 2019). The resulting embeddings are projected into a lower-dimensional space using UMAP (McInnes et al., 2020) to reveal clusters corresponding to recurring error types.

Each cluster is then summarized using GPT-5 (`gpt-5-2025-08-07`) to produce an interpretable description of the underlying error pattern. These summaries are used to define targeted error injections. For example, we observe a recurrent pattern where football/soccer clubs are mislabeled as CITY entities despite clear organizational context.

Errors from Model Test Logs. Finally, we mine actual errors from testing logs of an existing BERT-based de-identifier. From these logs, we extract incorrect entity mappings and construct a vocabulary of realistic errors, such as labeling *diabetes* as a HOSPITAL, *pleural* as a LOCATION, or *general* as a PROFESSION outside of a military context. These errors are injected verbatim, preserving their original surface forms and entity classes.

Error Dataset	Caught by Baseline ↑	Caught by Agent ↑
CARMEN-ES	2,714 / 2,807 (96.69%)	**2,779 / 2,807** (99.01%)
Transcripts-DE	248 / 294 (84.26%)	**292 / 294** (99.32%)

Table 2: Using the two medical error datasets, we evaluate both the baseline refiner and the agentic refiner on their abilities to detect and correct injected errors. The agentic refiner identifies and corrects over 99% of injected errors, clearly surpassing the baseline.

6 Discussion & Conclusion

We introduce an agent-based post hoc refinement layer for de-identification that improves precision of existing annotations by combining external memory, validation tools, adaptive context retrieval, and on-demand review skills.

In practice, excessive redaction can meaningfully degrade text quality. We observed frequent upstream errors involving disease eponyms, gene names, and institutional acronyms being labeled as PHI. Although conservative tagging improves privacy protection, indiscriminate removal can alter clinical meaning or remove medically salient

context. Agentic refinement allows more selective correction of such false positives, helping maintain alignment between de-identified text and its original informational content. This is particularly important when de-identified data are used for downstream modeling or large-scale pretraining.

The agentic pipeline was developed incrementally, with components introduced one at a time. In internal testing, each module yielded modest but consistent improvements in stability or performance. Among them, the modular review skills contributed most to precision gains. These skills encode detailed, class-specific guidelines and edge cases aligned with the annotation protocol used to construct our evaluation sets, enabling more consistent reasoning than a single prompt with context. Persistent memory (implemented as a to-do list) improves completeness and prevents skipped spans during multi-step review which can be a pattern common in contemporary agentic systems. Adaptive context retrieval primarily reduces token usage and unnecessary context, improving efficiency rather than directly affecting accuracy. This modular design allows practitioners to selectively enable components depending on deployment constraints or cost considerations. To contextualize these improvements, Appendix D includes a brief efficiency analysis of the proposed framework, reporting runtime, token usage, number of LLM calls, and approximate inference cost per document. Although the agentic pipeline introduces moderate additional computational overhead relative to the baseline LLM refiner, the increase remains bounded and corresponds to measurable gains.

Ultimately, the agent's improvements are not a trivial exchange of recall for precision. While some recall reduction is expected when removing spurious annotations, the agent consistently yields net F1 gains in overall annotation quality and utility, resolving false positives more selectively than baseline refinement. These results suggest that agentic refinement is a practical mechanism for precision control in high-recall de-identification pipelines.

Limitations

Currently, all experiments rely on a single LLM backend (GPT-5) with fixed inference parameters to reduce variability. While the proposed agentic framework is model-agnostic, performance characteristics may differ with other proprietary or open-weight models, particularly those with more limited reasoning capabilities. Preliminary experiments with the open-weight `gpt-oss-120b` model suggest that the framework generalizes beyond proprietary backbones, consistently improving precision, especially on transcript datasets, while exhibiting some recall trade-offs due to lower-recall upstream annotations in this setting.

Additionally, the agent operates strictly as a post hoc refiner and does not introduce new PHI/PII annotations. As a result, it cannot recover entities missed by the upstream de-identification system, and its recall is inherently bounded by the initial annotation set. Furthermore, although we evaluate across multiple languages and data modalities, broader experimentation across additional institutions, document types, and deployment settings would further strengthen generalizability.

The synthetic error dataset is designed to cover a wide range of common and systemic error types, but it does not exhaust the space of possible de-identification failures. While the datasets enable precise and controlled evaluation, they are intended to be extensible and can be expanded as new error patterns are observed in real-world use.

Finally, the agent's capabilities were implemented incrementally. Persistent task tracking, adaptive context windows, validation tools, and review skills were sequentially added as each module yielded observable performance improvements. A formal ablation study would be valuable for quantifying the isolated contributions of each component.

Acknowledgments

We gratefully acknowledge support from Paul Vozila from Microsoft during manuscript writing. JX acknowledges joint funding from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). This project was also supported by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with the UK Health Security Agency (UKHSA) (NIHR207397) and the NIHR Biomedical Research Centre, Oxford. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or UKHSA.

References

Kiana Aghakasiri, Noopur Zambare, JoAnn Thai, Carrie Ye, Mayur Mehta, J. Ross Mitchell, and Mohamed Abdalla. 2025. Not what the doctor ordered: Surveying LLM-based de-identification and quantifying clinical information loss. *Preprint*, arxiv:2509.14464 [cs].

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *Preprint*, arxiv:1904.03323 [cs].

Hong-Jie Dai, Tatheer Hussain Mir, Ching-Tai Chen, Chien-Chang Chen, Hao-Ping Yang, Chung-Hong Lee, Yi-Yun Chou, Yu-Chin Teng, Shalini Gupta, Omkar Panchal, Divyabharathy Ramesh Nadar, Wei-Hsiang Liao, Yu-Chuan Lin, Zi-Rui Zhao, Richard Tzong-Han Tsai, Yung-Chun Chang, and Jitendra Jonnagaddala. 2025. Leveraging large language models for the deidentification and temporal normalization of sensitive health information in electronic health records. *npj Digital Medicine*, 8(1):517. Publisher: Nature Publishing Group.

Zhuojun Ding, Wei Wei, and Chenghao Fan. 2025. Selecting and merging: Towards adaptable and scalable named entity recognition with large language models. *Preprint*, arxiv:2506.22813 [cs].

Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Marti, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. 2024. CARMEN-i: A resource of anonymized electronic health records in spanish and catalan for training and testing NLP tools.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. *Preprint*, arxiv:2211.10435 [cs].

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (MCP): Landscape, security threats, and future research directions. *Preprint*, arxiv:2503.23278 [cs].

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *Preprint*, arxiv:1904.05342 [cs].

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 214–221. Association for Computing Machinery.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025. Clinical ModernBERT: An efficient and long context encoder for biomedical text. *Preprint*, arxiv:2504.03964 [cs].

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. *Preprint*, arxiv:2311.08089 [cs].

Xinkui Lin, Yuhui Zhang, Yongxiu Xu, Kun Huang, Hongzhang Mu, Yubin Wang, Gaopeng Gou, Li Qian, Li Peng, Wei Liu, Jian Luan, and Hongbo Xu. 2025. MAKAR: a multi-agent framework based knowledge-augmented reasoning for grounded multimodal named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6121–6141. Association for Computational Linguistics.

Meng Lu, Yuzhang Xie, Zhenyu Bi, Shuxiang Cao, and Xuan Wang. 2025. CROSSAGENTIE: Cross-type and cross-task multi-agent LLM collaboration for zero-shot information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13953–13977. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arxiv:1802.03426 [stat].

Ishna Neamatullah, Margaret M. Douglass, Li-wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8:32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Preprint*, arxiv:1908.10084 [cs].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arxiv:2302.04761 [cs].

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face. *Preprint*, arxiv:2303.17580 [cs].

Praphul Singh, Charlotte Dzialo, Jangwon Kim, Sumana Srivatsa, Irfan Bulu, Sri Gadde, and Krishnaram Kenthapadi. 2025. RedactOR: An LLM-powered framework for automatic clinical data de-identification. *Preprint*, arxiv:2505.18380 [cs].

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. BioClinical ModernBERT: A state-of-the-art long-context encoder for biomedical and clinical NLP. *Preprint*, arxiv:2506.10896 [cs].

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58 Suppl:S20–S29.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533. Association for Computational Linguistics.

Ozlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018. Describing a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 10–21, Tilburg University, The Netherlands. Association for Computational Linguistics.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4431–4441. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. *Preprint*, arxiv:2210.03629 [cs].

Barry Zhang, Keith Lazuka, and Mahesh Murag. 2025. Equipping agents for the real world with agent skills. Published October 16, 2025. Updated December 18, 2025 to announce Agent Skills as an open standard.

Guangcheng Zhu, Ruixuan Xiao, Haobo Wang, Zhen Zhu, Gengyu Lyu, and Junbo Zhao. 2025. Large margin representation learning for robust cross-lingual named entity recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4270–4291. Association for Computational Linguistics.

A Prompts

A.1 Original Annotation Prompt

The following prompt was used to generate the initial PHI/PII annotations on which all refinement experiments were conducted:

GPT-5

```
You will be provided a note containing named entities.
Please identify all instances of the named entity: {ENTITY}.

Specific annotation instructions for the entity: {ENTITY_INSTRUCTION}

Respond with the entity type and the text of the entity, with one entity per line.
Here is an example for {LANGUAGE}: {EXAMPLE}

Respond with the text from the note EXACTLY.

...

If no annotations are found, respond with NO ENTITIES.

The user will provide the note to annotate.
```

A.2 Baseline Refinement Prompt

The following prompt was used to perform baseline LLM-based refinement of PHI/PII annotations:

GPT-5

```
You are correcting named entity recognition annotations.

You will be provided with a single named entity annotation and the context.

You must respond with one of the following:
    * correct - The provided annotation is correct.
    * update - There is an entity, but it must be corrected in some way.
    * remove - The annotated text does not correspond to the given entity type.

{ENTITY GUIDELINES}

{EXAMPLES}

The user will provide the annotation to correct.
```

B Example Agentic Output

The following examples illustrate representative cases where the single-pass LLM refiner failed to appropriately remove incorrect PHI/PII annotations, either due to insufficient contextual reasoning or inconsistent application of labeling guidelines. In contrast, the agentic refiner successfully identifies and corrects these errors by combining persistent task tracking, targeted context retrieval, deterministic validation, and on-demand review skills. These examples highlight how agentic refinement enables more reliable and selective correction of false positives than prompting alone.

Example 1 (Mentions of Medication Manufacturers). This example illustrates systematic over-tagging of pharmaceutical manufacturers and brand names as ORGANIZATION entities in a therapeutic plan. Although the spans correspond to real companies, the surrounding context makes clear that they function purely as medication descriptors and do not identify patients, providers, or institutions. By loading the organization-specific review skill and retrieving targeted context for each span, the agent consistently recognizes these mentions as non-PHI and removes all annotations. This case highlights the agent's ability to apply contextual guidelines at scale and eliminate large batches of false positives that frequently occur in medication lists.

Original annotations (CARMEN-I PLAN TERAPEUTICO):

Note ID	Index ($s_i - e_i$)	Span	Entity Class
<NOTE_ID>,	29-35,	normon,	ORGANIZATION
<NOTE_ID>,	128-139,	kern pharma,	ORGANIZATION
<NOTE_ID>,	223-228,	davur,	ORGANIZATION
<NOTE_ID>,	333-338,	mylan,	ORGANIZATION
<NOTE_ID>,	685-696,	kern pharma,	ORGANIZATION
<NOTE_ID>,	782-790,	teofarma,	ORGANIZATION
<NOTE_ID>,	912-916,	teva,	ORGANIZATION
<NOTE_ID>,	1181-1186,	cinfa,	ORGANIZATION
<NOTE_ID>,	1279-1284,	cinfa,	ORGANIZATION
<NOTE_ID>,	1380-1387,	tevagen,	ORGANIZATION
<NOTE_ID>,	1606-1611,	pensa,	ORGANIZATION
<NOTE_ID>,	1718-1724,	apotex,	ORGANIZATION

Agentic Refiner

Agent Summary:

- Annotations Reviewed: 12
- Annotations Modified: 0
- Annotations Deleted: 12

Actions Taken:

- Loaded the ORGANIZATIONS skill. Per guidelines, vaccine/pill/device brands and pharmaceutical manufacturer names (e.g., Normon, Kern Pharma, Mylan, Teva, Cinfa, etc.) should NOT be tagged as ORGANIZATION in medication lists, as they are not PHI in this context.
- Retrieved context for each item and confirmed all were medication manufacturer/brand mentions in a therapeutic plan.
- Deleted all 12 ORGANIZATION annotations accordingly.
- Final validation returned no errors. The final annotation list is empty.

Major Issues Resolved: Removal of 12 non-PHI/incorrectly labeled items (organizations that are not PHI in context).

Example 2 (Mixed Clinical Backgrounds). This example demonstrates selective refinement in a document containing a diverse set of entity types, including dates, locations, organizations, and clinical terminology. The agent removes only two incorrectly labeled spans: an ORGANIZATION tag applied to a disease classification token (*GOLD*) and a LOCATION tag applied to an anatomical term (*cisura silviana*). All other annotations are retained after context-aware verification, including correctly labeled hospitals, clinicians, dates, and ages. This example shows that the agent does not indiscriminately trade recall for precision, but instead performs targeted corrections while preserving valid PHI annotations.

Original annotations (CARMEN-I IA ANTECEDENTES):

Note ID	Index ($s_i - e_i$)	Span	Entity Class
...			
<NOTE_ID>,	2453-2457,	GOLD,	ORGANIZATION
...			
<NOTE_ID>,	2918-2933,	cisura silviana,	LOCATION

Agentic Refiner

Agent Summary:

- Annotations Reviewed: 28
- Annotations Modified: 0
- Annotations Deleted: 2

Actions Taken:

- Loaded relevant skills: REVIEW, DATES-AGES, LOCATIONS, NAMES-PEOPLE, PROFESSIONS, ORGANIZATIONS.
- Checked rules for DATE entries; none qualified for automatic acceptance; proceeded with manual review.
- Retrieved surrounding context for all annotations.
- Manually reviewed problematic spans:
  - Deleted ORGANIZATION 'GOLD' - context shows this as a COPD severity/classification term (non-PHI).
  - Deleted LOCATION 'cisura silviana' - identified as an anatomical/fissure term (non-PHI).
- Verified remaining annotations:
  - Ages and professions correctly labeled and contextually PHI.
  - Hospital `C. La Salud` and doctor `Dr. Ager` correctly labeled and complete.
  - Dates (years, month-year, numeric formats, weekdays, and schedule abbreviations `Ma/Ju/Sa`) are tied to patient care timeline; boundaries and labels confirmed.
- Final validation succeeded with no errors or warnings. Remaining annotations marked as correct per guidelines.

Major Issues Resolved: Removal of 2 non-PHI/incorrectly labeled items (classification token and anatomical label).

C Regular Expression Rules for Error Injection

We inject a set of deterministic, regular expression-driven error patterns to simulate common annotation mistakes. Each rule targets a realistic failure mode observed in deployed de-identification systems or LLM annotation outputs. The 10 rules used in our experiments are described below.

Ambiguous common-word names. Many surnames or given names are also common English words (*e.g.*, `Brown`, `May`, `Grant`, or `Jordan`). This rule treats such tokens as NAME/PATIENT entities in contexts where they are not PHI (*e.g.*, "brown" in "the patient reported a brown rash" → incorrectly labeled as PATIENT).

Example

"Brown" ⇒ injected as NAME.
"May" in "the patient may require surgery" ⇒ injected as NAME.
"Grant" in "we will grant discharge tomorrow" ⇒ injected as PATIENT.
"Mann" in "der mann klagt über Schmerzen" ⇒ injected as PATIENT.

Two-digit lab values mislabeled as AGE. Numeric laboratory or microbiology results (often two digits) appear near medical keywords and can be mistaken for patient ages. We identify two-digit numbers near medical tokens and relabel them as AGE.

Example

"12" in "WBC 12" near "culture" ⇒ injected as AGE.
"45" in "platelets 45" near "CBC" ⇒ injected as AGE.
"22" in "RR 22" ⇒ injected as AGE.
"45" in "Leukozyten 45" nahe "Infektion" ⇒ injected as AGE.

Blood-pressure values labeled as DATE. Blood-pressure strings such as "120/80" resemble date-like patterns and may be mis-tagged as DATE. We search for paired numeric tokens with a slash in clinical contexts and relabel them as DATE.

Example

"120/80" in "BP 120/80" ⇒ injected as DATE.
"130/85" in "BP 130/85 sitting" ⇒ injected as DATE.
"90/60" in "hypotensive at 90/60" ⇒ injected as DATE.
"130/85" in "Blutdruck 130/85 gemessen" ⇒ injected as DATE.

City → State/Organization confusion. Place-name granularity errors: city names are relabeled as either broader administrative units or organizations. This simulates errors where classifiers confuse geographic granularity or entity class.

Example

"Berlin" ⇒ injected as STATE.
"Madrid" in "ingresado en Madrid ayer" ⇒ injected as ORGANIZATION.
"Paris" in "treated in Paris clinic" ⇒ injected as ORGANIZATION.

Generic healthcare words → HOSPITAL. Terms that denote wards, units, or specialties (*e.g.*, `ICU`, `cardiology`, or `ward`) are relabeled as full HOSPITAL entities. This models over-generalization errors.

Example

"ICU" ⇒ injected as HOSPITAL.
"cardiology" in "admitted to cardiology" ⇒ injected as HOSPITAL.
"ER" in "sent to ER for evaluation" ⇒ injected as HOSPITAL.

Patient/Doctor confusion. Swap PATIENT and DOCTOR labels to simulate role-disambiguation mistakes (*e.g.*, doctor names annotated as patients and vice versa).

Example

"Dr. Smith" ⇒ injected as PATIENT.
"John Davis" in "Dr. John Davis reviewed labs" ⇒ injected as PATIENT.

Age/Date confusion. Swap AGE and DATE labels to mimic numeric ambiguity (*e.g.*, two- or four-digit tokens interpreted as a birth year vs. age).

Example

"1978" in "Born 1978" ⇒ injected as AGE.
"84" in "patient is 84" injected as DATE.
"63" in "63-jähriger Patient" ⇒ injected as DATE.

Hospital/Organization confusion. Map specific hospital names to a generic ORGANIZATION label intended for employers/companies, simulating coarse-grained classification errors.

Example

"Hospital Clínic" ⇒ injected as ORGANIZATION.
"Massachusetts General Hospital" ⇒ injected as ORGANIZATION.
"Universitätsklinikum Heidelberg" ⇒ injected as ORGANIZATION.

ZIP / City / State swaps. Exchange geographic and postal labels (ZIP, city, or state) to reflect location-granularity errors and numeric/non-numeric confusions.

Example

"02115" (ZIP) ⇒ injected as CITY.
"CA" ⇒ injected as CITY.

Eponymous diseases as PATIENT. Diseases named after people (*e.g.*, `Parkinson's` or `Alzheimer's`) are injected as PATIENT entities, modeling a frequent confusion where models treat eponyms as person names.

Example

"Parkinson" in "Parkinson's disease" ⇒ injected as PATIENT.
"Hodgkin" in "Hodgkin lymphoma" ⇒ injected as PATIENT.
"Morbus Alzheimer" ⇒ injected as PATIENT.

Procedures. Each rule is applied deterministically over reference annotations or raw text according to a priority order to avoid conflicting injections (*e.g.*, numeric-format rules are applied before generic city/state swaps). Injected errors are recorded with provenance (rule id, original label, injected label) to enable precise evaluation when using the error dataset.

These rules are designed to be realistic but conservative: they target representative failure modes rather than exhaustively enumerating every possible error. The rule set is extensible; new patterns discovered in deployment logs can be added to the vocabulary and re-run to produce larger or more diverse synthetic error sets.

D Comparison of Computational Costs

Metric	LLM Baseline	Agentic Refinement
# LLM Calls	420	161
Time to Process (seconds)	1,124	1,820
Input Tokens	2,284,378	3,076,333
Output Tokens	432,270	720,450
Approx. Cost (USD)	$7.18	$11.05

Table 3: Efficiency comparison between the single-pass LLM refinement baseline and the proposed agentic refinement pipeline on a subset of the CARMEN-ES dataset (3,314 tokens; 84 annotations). Wallclock time reflects end-to-end single-threaded execution, including preprocessing and postprocessing. Costs are approximate and based on standard GPT-5 pricing.

Do Syntactic Features Help Biomedical Relation Extraction? An Empirical Study of Verb Token and Dependency Graph Augmentation

Mustafa Kamal Sikder and Ernest Kwegyir-Afful
Hazard Assessment and Analytics Branch
Human Foods Program
U.S. Food and Drug Administration
{Abu.Sikder, Ernest.Kwegyir-Afful}@fda.hhs.gov

Abstract

We investigate whether explicit syntactic features improve transformer-based biomedical relation extraction when added to typed entity marker pooling. We evaluate two augmentation strategies on top of BiomedBERT: (1) verb token augmentation, which concatenates the hidden state of the dependency root verb to the entity representations, and (2) a two-layer graph convolutional network (GCN) that refines encoder hidden states over the dependency parse before entity pooling. We experimented on three biomedical datasets: ChemProt, DDI, and AIMed with three random seeds. We found neither strategy consistently outperformed the entity-only baseline. The GCN yielded modest gains on AIMed ($+0.007$ F1) and ChemProt ($+0.003$ F1) but decreased performance on DDI (-0.013 F1). Verb token augmentation helps only on AIMed ($+0.004$ F1) and underperforms on the other two datasets. A syntactic characterization of the datasets reveals that DDI has substantially higher passive voice usage (50.7% of relation-bearing sentences) than AIMed (27.0%) or ChemProt (30.9%), suggesting that syntactic augmentation is more effective when sentences exhibit active verbal structure with semantically informative predicates. These results suggest that corpus-level syntactic characteristics, particularly passive voice usage, may moderate the utility of explicit syntactic augmentation, though the small magnitude of observed differences warrants caution in interpretation.

1 Introduction

Relation extraction (RE) is a core task in biomedical natural language processing that involves identifying semantic relationships between entities in text (Madan et al., 2019; Huang et al., 2024). Automatically detecting chemical-protein, drug-drug, and protein-protein interactions from scientific literature supports downstream applications in pharmacovigilance, knowledge base construction, and public health and food safety surveillance, including hazard signal detection, adverse event monitoring, and contamination source identification. (Zhang et al., 2025; Bawankar and Kumar, 2026; Munkhdalai et al., 2018; Zuo et al., 2022).

The dominant approach to RE with pretrained language models involves inserting typed entity markers around entity spans in the input sequence, and the model is trained to classify relations from the pooled representations at marker positions (Zhong and Chen, 2021; Ye et al., 2022). This approach has been widely adopted for biomedical RE due to its simplicity and strong empirical performance (Gu et al., 2021).

In most biomedical sentences, the relational predicate or the verb connecting two entities carries essential information about the nature of the relationship. For example, in "aspirin inhibits COX-2", the verb inhibits is the primary signal distinguishing an inhibitor relation from an activator or substrate relation. Majewska et al. (2021) demonstrated that biomedical verbs exhibit domain-specific semantic-syntactic properties distinct from general English, and that structured verb knowledge can improve biomedical NLP models, motivating the hypothesis that the predicate verb carries informative relational signal in biomedical sentences. Additionally, prior work on syntax-augmented RE has demonstrated that incorporating dependency structure can improve performance, particularly for long-range relations (Zhang et al., 2018). However, these approaches were not applied on top of the token representations of the language model encoders.

In this paper, we investigate two strategies for incorporating syntactic information into entity marker-based RE: (1) verb token augmentation, which concatenates the hidden state of the main predicate verb to the entity representations; and (2) dependency graph convolutional network (GCN), which refines entity representations through graph

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 128–134
July 3-4, 2026 ©2026 Association for Computational Linguistics

convolution over the dependency parse before classification. We evaluated both strategies against an entity-only baseline across three biomedical datasets: ChemProt (Krallinger, 2017), the DDI corpus (Herrero-Zazo et al., 2013), and AIMed (Bunescu et al., 2005), using BiomedBERT (Gu et al., 2021) as the base encoder.

2 Related Works

The entity marker approach for relation extraction was introduced by Soares et al., 2019, who inserted special boundary tokens around entities and used their hidden states for classification. Zhong and Chen (2021) showed that typed markers encoding entity categories (e.g., <S:CHEMICAL>) significantly outperform generic ones, and Ye et al. (2022) proposed packed levitated markers that avoid disrupting the original token sequence. Zhong and Chen (2021) demonstrated that concatenating entity start-marker representations with the sentence-level [CLS] embedding yields strong performance with minimal architectural overhead — an approach we adopt as our baseline. In the biomedical domain, BioBERT (Lee et al., 2019) and BiomedBERT (Gu et al., 2021) established that domain-specific pretraining combined with entity markers achieves competitive results on benchmarks such as ChemProt and DDI.

Our work builds on two lines of syntactic augmentation. Zhang et al. (2018) showed that graph convolution over pruned dependency trees improves relation extraction, particularly for long-range entity pairs, though their method operated on static embeddings rather than transformer hidden states. Majewska et al. (2021) demonstrated that biomedical verbs exhibit domain-specific knowledge and retrofitting that into the model can improve classification tasks. We extend both ideas to the transformer entity-marker setting: verb token augmentation directly concatenates the predicate's contextual representation, while our dependency GCN applies graph convolution over the full parse tree on top of BiomedBERT's hidden states.

3 Methods

We investigate three relation representation strategies built on top of a shared transformer encoder. All three share the same input format, training procedure, and evaluation protocol, differing only in how the final relation representation is constructed before classification.

3.1 Base Encoder and Input Representation

We use BiomedBERT (Gu et al., 2021) as our encoder across all conditions. For each candidate entity pair in a sentence, we insert typed markers around the subject and object spans following the PURE framework (Zhong and Chen, 2021).

$$[CLS] < S_S E_1 > aspirin < S_E E_1 > ... \\ < O_S E_2 > COX - 2 < O_E E_2 > [SEP]$$

Where $[CLS]$ and $[SEP]$ are encoder's special tokens, S_S and S_E indicate the start and end of the subject entity markers and O_S and and O_E indicate the object entity marker, and E_1, E_2 were the entity categories (e.g., CHEMICAL, GENE-Y, DRUG). Markers are typed with the entity category so the encoder can distinguish entities semantically. These marker tokens were added to the tokenizer vocabulary and their embeddings are initialized to the vocabulary mean to avoid out-of-distribution activations at the start of fine-tuning. One training example is generated per ordered entity pair per sentence; pairs with no gold relation are assigned the $no_relation$ label.

The encoder processes the marked-up sequence and produces contextual hidden states $h \in \mathbb{R}_{seq_len \times 768}$. All three classification models described below operate on these hidden states.

3.2 Entity Marker Pooling (Baseline)

The updated entity representations were obtained from the encoder last hidden state at the position of each entity start marker (e.g. $S_S E_1$ and $O_S E_2$). The vectors were concatenated and passed through layer normalization, dropout, and a linear classifier to produce logits over relation types. This served as the reference condition for all comparisons.

3.3 Verb Token Augmentation

For this condition, together with the entity marker embeddings, we obtained the representation of the main predicate verb and concatenated that with the entity markers. We used spaCy's dependency parser (Honnibal et al., 2020) with $en_core_web_sm$ model. We selected the ROOT token of the dependency tree if its part-of-speech tag was VERB or AUX. If a sentence contained multiple ROOT candidates, we selected the one closest to the subject entity span center. When no ROOT verb was found, a special NO_VERB_TOKEN embedding was appended, initialized to the vocabulary mean.

Unlike the entity markers, which point to dedicated special tokens, the verb index points to the first subword token of the verb word itself, preserving the lexical identity of the predicate alongside its contextual representation. Thus, the hidden state encoded the verb's semantic and syntactic context.

The verb index is extracted from the original (un-marked) sentence token sequence and then adjusted to account for the marker tokens inserted during feature creation. Verb extraction is performed once per sentence during preprocessing and cached for all entity pairs sharing the same sentence. The NO_VERB_TOKEN fallback was triggered for 10.3%, 11.7%, and 7.8% of relations in ChemProt, AIMed, and DDI respectively, indicating that a meaningful minority of instances received no relational verb signal regardless of dataset.

3.4 Dependency Graph Neural Network

In the entity + graph convolutional network (GCN) condition, we replace the direct pooling step with a two-layer graph convolutional network (Kipf and Welling, 2016) that refines the model's hidden states according to the syntactic dependency structure of the sentence before extracting entity representations.

Word-level pooling: Because spaCy operates at the word level while BiomedBERT uses WordPiece subtokenisation, we first mean-pool the BiomedBERT hidden states of all subtokens belonging to the same original word. Marker tokens ($< S_S E_1 >$, etc.) are excluded from this alignment — they have no corresponding spaCy word. This produces a word-level embedding matrix $X \in \mathbb{R}_W \times 768$ where W is the number of original words (capped at 128).

Dependency graph construction: We ran dependency parser on the original unmarked sentence to obtain a dependency parse. We built an undirected adjacency matrix $A \in \mathbb{R}_W \times W$ where $A[i, j] = 1$ if word i and word j share a head-child dependency relation, and $A[i, i] = 1$ (self loops). The matrix is symmetrically normalised as $D^{-1/2}AD^{-1/2}$ where D is the diagonal degree matrix. Adjacency matrices are computed once per unique sentence and cached across all entity pairs from the same sentence.

GCN message passing: The word embeddings are refined through two GCN layers:

$$X^1 = \text{ReLU}(A_normX^0W^1)... \\ [W \times 768 \rightarrow W \times 256]$$

$$X^2 = \text{ReLU}(A_normX^1W^2)... \\ [W \times 256 \rightarrow W \times 768]$$

Each layer aggregates information from syntactically connected words. After two layers, each word's representation reflects information from words up to two dependency hops away. Because the dependency graph is undirected, information flows both from head to child and from child to head. The verb node aggregates from its subject and object arguments, and the entity nodes aggregate from the verbs governing them.

Classification: The GCN-refined embeddings at the subject and object word positions are extracted and passed to the classifier:

$$r = LayerNorm([X^2[s_word]; X^2[o_word]])$$

Note that s_word and o_word are the word-level indices of the first token of each entity span in the original sentence. The classifier was the same linear layer used in the other conditions, with input dimension 2×768. The GCN weights are randomly initialized and trained jointly with the BiomedBERT encoder from scratch during fine-tuning.

3.5 Training

All three conditions share identical training hyperparameters. We fine-tuned with AdamW (weight decay 0.01 on non-bias/LayerNorm parameters), a linear learning rate schedule with 10% warmup, batch size 16, and learning rate 2×10^{-5}. The maximum sequence length is 256 for ChemProt and AIMed and 384 for DDI (which contains longer sentences). We train for 10 epochs on AIMed and 5 epochs on ChemProt and DDI, selecting checkpoints by the development set F1. All experiments are repeated with three random seeds (0, 1, 42). Our primary evaluation metric was micro F1 over positive relation classes. We reported the mean F1 for all seeds with standard deviation. The loss of Cross-entropy is calculated over all pairs of candidate entities, with $no_relation$ as a negative class. Gradient clipping is applied at norm 1.0. Marker token embeddings and GCN weights are initialized to the vocabulary mean and random normal respectively.

3.6 Datasets

We evaluate on three biomedical relation extraction datasets spanning chemical-protein, drug-drug, and protein-protein interaction tasks. All datasets used the entity markers described in 3.1 with entity type CHEMICAL, DRUG, or PROTEIN as appropriate.

ChemProt ChemProt (Krallinger, 2017) is a sentence-level relation extraction dataset from the BioCreative VI challenge covering interactions between chemicals and gene/protein targets in biomedical abstracts. Each candidate pair is classified into one of five relation groups (CPR:3 upregulator, CPR:4 downregulator, CPR:5 agonist, CPR:6 antagonist, CPR:9 substrate) or assigned no relation. We use the official BioCreative VI train, development, and test partitions and evaluate using micro F1 over the five positive CPR groups.

DDI Corpus The DDI corpus (Herrero-Zazo et al., 2013) contains pharmacological sentences annotated for drug-drug interactions, drawn from DrugBank database entries and MedLine abstracts. Relations are classified into four types: MECHANISM (pharmacokinetic mechanism of interaction), EFFECT (clinical effect of the interaction), ADVISE (recommendation against co-administration), and INT (unspecified interaction). We use the SemEval 2013 Task 9 train and test partitions. We hold out 20% of the training data for development and use the remaining 80% for training. We evaluate using micro F1 over the four positive relation types.

AIMed AIMed (Bunescu et al., 2005) is a binary protein-protein interaction dataset drawn from MEDLINE abstracts. Each candidate entity pair is labelled as interacting (INTERACT) or not. We use fold 1 of the original data and hold out 10% of the training set as development set using a fixed random seed. While results may vary across folds, the fold assignment only affects the train/dev/test split rather than the underlying data

4 Results

We present results across three conditions: entity marker pooling (Entity), verb token augmentation (Entity+Verb), and dependency GCN (GCN) on three datasets: ChemProt, DDI, and AIMed. All are sentence-level datasets.

4.1 Syntactic Characterization of Datasets

To contextualize the model results, we characterize the syntactic structure of each dataset's test set using automatic dependency parsing (spaCy $en_core_web_sm$). We analyzed all sentences containing at least one gold relation. Table 1 presents sentence-level and relation-level syntactic metrics across the three datasets.

Dependency tree depth: The average tree depth is 7.5, 7.5, and 7.1 for AIMed, ChemProt, and DDI respectively. This indicates that the three datasets have similar structural complexity as measured by tree depth. The datasets also have similar average sentence lengths (26–31 tokens).

Passive voice usage: Passive constructions appear in 50.7% of DDI sentences containing relations, compared to 27.0% in AIMed and 30.9% in ChemProt. At the relation level, 33.6% of DDI gold relations occur in sentences where the ROOT verb is a passive construction which is higher than ChemProt (20.2%) and AIMed (18.9%). The passive ROOT verb is an unreliable predicate signal for relation extraction: in a passive sentence such as "Synergism was observed between methylglyoxal and piperacillin", the ROOT verb observed describes the research act rather than the pharmacological interaction between the entities.

4.2 Model Performance

Table 2 shows test F1 ($mean \pm std$) across three seeds) for all three conditions on all three datasets. Full per-seed results are provided in the Appendix.

Verb token augmentation helped AIMed dataset and underperformed the entity-only baseline on ChemProt and DDI datasets. The GCN condition helped AIMed and ChemProt datasets and performed better than entity+verb condition in the DDI dataset. On DDI, neither augmentation strategy improves over the entity-only baseline. On ChemProt, GCN marginally outperforms entity-only ($+0.003$) method.

Seed-level data suggested additional patterns. On DDI, both augmentation strategies underperformed the baseline across all three seeds, indicating systematic incompatibility. Verb augmentation also substantially increased training instability on DDI (F1 range: $0.755 - 0.770$) compared to the baseline ($0.778 - 0.784$). In contrast, the GCN outperformed the baseline in all three seeds on AIMed. On ChemProt, the GCN matched or exceeded the baseline in all three seeds despite small mean differences, suggesting a consistent modest advantage. To assess whether the observed differences are reliable, we ran paired approximate randomization tests comparing each augmentation strategy against

Metric	AIMed	ChemProt	DDI
Avg sentence length (tokens)	26.2 ± 12.7	30.5 ± 14.6	26.4 ± 13.5
Avg dependency tree depth	7.5 ± 2.1	7.5 ± 2.2	7.1 ± 2.3
Avg dep path between entities (hops)	4.8 ± 2.8	5.6 ± 2.6	6.6 ± 3.3
% sentences with passive voice	27.0%	30.9%	50.7%
% sentences with active ROOT verb	58.6%	55.3%	45.0%
% relations: active verb between entities	11.7%	25.5%	28.4%
% relations: passive ROOT verb	18.9%	20.2%	33.6%
% relations: copular ROOT verb	10.8%	13.3%	9.1%

Table 1: Syntactic characterization of test sets. Sentence-level metrics computed over sentences containing at least one gold relation. Dependency path lengths computed between subject and object entity head tokens. Parser: spaCy *en_core_web_sm*.

Condition	AIMed	ChemProt	DDI
N	222	3463	964
Entity (baseline)	0.818 ± 0.002	0.760 ± 0.006	$\mathbf{0.780 \pm 0.003}$
Entity + Verb	0.822 ± 0.002	0.755 ± 0.006	0.762 ± 0.006
Entity + GCN	$\mathbf{0.825 \pm 0.004}$	$\mathbf{0.763 \pm 0.003}$	0.767 ± 0.006

Table 2: Mean test $F1 \pm std$ across three seeds $(0, 1, 42)$. Bold marks best per dataset. Micro F1 over positive relation classes only. $N =$ number of positive test instances

the baseline separately per dataset (full results in Appendix). No comparison reached statistical significance (all $p > 0.05$), consistent with the small effect sizes and the limited statistical power of three seeds.

5 Discussions

5.1 GCN Outperforms Baseline Where Active Verbal Structure is More Prevalent

The dependency GCN improved over the entity-only baseline on two of three datasets: AIMed ($+0.007$) and ChemProt ($+0.003$). On DDI it underperformed (-0.013). This pattern aligns with a key difference in sentence structure across datasets. AIMed has the lowest passive construction rate (27.0%) and a relatively high proportion of sentences with an active ROOT verb (58.6%). ChemProt is similar (30.9% passive, 55.3% active ROOT). DDI, in contrast, has a substantially higher passive rate (50.7%) and fewer sentences with an active ROOT verb (45.0%). A likely explanation of these could be: when sentences followed an active structure, the dependency graph provided paths between entities that passed through semantically informative verbs. GCN message passing propagated this predicate information into the entity representations before classification. When sentences were predominantly passive, the ROOT verb was a reporting predicate rather than the relational one, and the dependency paths between entities carried less useful signal. This interpretation is consistent with the direction of results across all three datasets, though we caution that three datasets is insufficient to establish this relationship quantitatively.

5.2 Verb Token Augmentation is Inconsistent

Verb token augmentation improves over the baseline only on AIMed ($+0.004$) and underperforms on ChemProt (-0.005) and DDI (-0.018). The inconsistency is likely attributable to the noise floor of automatic ROOT verb extraction. For only 11.7% of AIMed relations, 25.5% of ChemProt relations, and 28.4% of DDI relations (Table 1), an active verb was directly between the two entity spans. Indicating, the smaller proportions of sentences where the ROOT verb reliably signaled the relation. In the remaining cases, the concatenated verb token added a representation that is structurally present but semantically peripheral, competing with rather than complementing the entity marker representations. That verb token helps on AIMed despite its low verb-between-entity rate (11.7%) is counterintuitive to this hypothesis. One possible explanation is that AIMed's binary task structure may be easier to benefit from any additional context, even imperfectly located verbal signal. A multi-class setting like ChemProt, where fine-grained relation types must be distinguished,

may penalize noisy verb tokens more heavily. This remains speculative without further controlled experiments.

5.3 Limitations and Future Directions

Our experiments have several limitations. All syntactic features are extracted using a general-domain dependency parser (spaCy $en_core_web_sm$) not adapted to biomedical text. Parser errors such as, correctly detecting noun phrases and passive construction on complex biomedical sentences can propagate into both the verb token and GCN conditions. A biomedical dependency parser trained on domain-specific treebanks may yield more reliable syntactic features and alter the results for both augmentation strategies. We evaluate on a single encoder (BiomedBERT) and three seeds. Different encoders or a larger seed set may produce different conclusions. For AIMed we use only fold 1 of the 10-fold cross-validation structure; results may not generalize across all folds. Additionally, the GCN improvements on AIMed and ChemProt are small (0.003–0.007) and within or near one standard deviation of the baseline, so they should not be overstated. In future work we will investigate biomedical-domain dependency parsers, pruned dependency paths (Zhang et al., 2018) rather than the full graph.

6 Conclusions

We presented a controlled empirical study of two syntactic augmentation strategies: verb token concatenation and dependency GCN, and compared to entity marker pooling for biomedical relation extraction. Experiments on ChemProt, DDI, and AIMed using BiomedBERT with three random seeds showed that neither strategy consistently outperformed the entity-only baseline. The dependency GCN improved on AIMed ($+0.007$) and ChemProt ($+0.003$) but underperformed on DDI (-0.013). Verb token augmentation helped only on AIMed ($+0.004$).

Our syntactic analysis showed that datasets with lower passive construction rates (AIMed 27.0%, ChemProt 30.9%) benefit from GCN augmentation, while the passive-heavy DDI corpus (50.7%) did not. This suggests that the utility of explicit syntactic features in transformer-based RE depends on the linguistic characteristics of the underlying corpus rather than being universally beneficial. These findings offer practical guidance for biomedical RE systems supporting public health surveillance. When working with literature exhibiting predominantly active sentence structures, dependency GCN augmentation may provide modest improvements. For corpora with high passive voice usage, such as DDI, entity marker pooling alone could be more reliable compared to either augmentation strategy. This is particularly relevant for public health and food safety applications that rely on extracting chemical interactions, adverse events, and contamination signals from biomedical literature, where understanding which augmentation strategies generalize across different writing styles can improve robustness and reliability. For real-world deployment in applications such as chemical hazard surveillance, practitioners can characterize the syntactic profile of their target corpus and validate augmentation strategies on representative data before deployment. We release all code and converted datasets to support reproducibility and encourage future work to explicitly report syntactic characteristics of biomedical RE benchmarks.[1]

References

Sapna Bawankar and Sanjay Kumar. 2026. Scalable knowledge extraction from biomedical literature for drug discovery applications. *Procedia Computer Science*, 275:226–233. 7th International Conference on AI in Computational Linguistics.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155. Information Extraction and Summarization from Medical Documents.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

[1] https://github.com/mustafasikder/Biomedical-Relation-Extraction

Ming-Siang Huang, Jen-Chieh Han, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2024. Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. *Briefings in Bioinformatics*, 25(3):bbae132.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Rabal-O. Lourenço A. Krallinger, M. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop,*, 141-146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sumit Madan, Justyna Szostak, Ravikumar Komandur Elayavilli, Richard Tzong-Han Tsai, Mehdi Ali, Longhua Qian, Majid Rastegar-Mojarad, Julia Hoeng, and Juliane Fluck. 2019. The extraction of complex relationships and their conversion to biological expression language (bel) overview of the biocreative vi (2017) bel track. *Database*, 2019.

Olga Majewska, Charlotte Collins, Simon Baker, Jari Björne, Susan Windisch Brown, Anna Korhonen, and Martha Palmer. 2021. Bioverbnet: a large semantic-syntactic classification of verbs in biomedicine. *Journal of Biomedical Semantics*, 12(1).

Tsendsuren Munkhdalai, Feifan Liu, and Hong Yu. 2018. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. *JMIR Public Health and Surveillance*, 4(2):e29.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *CoRR*, abs/1906.03158.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, Jian Wang, Donghu Sun, Henry Chung, Jun Zhou, Eric Zhou, Ben Lee, Peili Zhang, Xing Qiu, Tingting Zhao, and Jinfeng Zhang. 2025. A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. *Nature Machine Intelligence*, 7(4):602–614.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Min Zuo, Baoyu Zhang, Qingchuan Zhang, Wenjing Yan, and Dongmei Ai. 2022. An entity relation extraction method for few-shot learning on the food health and safety domain. *Computational Intelligence and Neuroscience*, 2022:1–11.

Beyond Knowledge Graphs: PubMedBERT Embeddings as a Competitive Standalone Modality for Drug Repurposing

Rishik K. Kondadadi
University of Minnesota
Eastview High School
konda052@umn.edu

John E. Ortega
Northeastern University
j.ortega@northeastern.edu

Abstract

Drug repurposing methods rely heavily on knowledge graph (KG) embeddings, but building and curating these graphs takes considerable effort. We present two findings on the Hetionet drug-disease benchmark and an epilepsy ranking task. *First*, PubMedBERT text embeddings, fed through the same downstream classifiers and identical 10-fold splits as four re-trained KG baselines (TransE, ComplEx, DistMult, RotatE), reach AUROC 0.910, above all four (best: RotatE, 0.854); a Random Forest on the same vectors scores 0.880. The comparison is asymmetric in one important way: PubMedBERT was pretrained on the literature Hetionet was curated from, so the result is best read as "text-with-literature-supervision vs. graph-only," and a head-to-head with text-augmented KG methods (KG-BERT, TxGNN) is left as follow-up. *Second*, across all seven combinations of text, molecular (ECFP4), and gene expression (LINCS L1000) features, cross-attention fusion of weaker modalities into text consistently degrades performance, despite a gated mechanism intended to suppress unhelpful modalities; the residual path forces the strong modality to absorb noise. The model also ranks proconvulsants (amoxapine, flumazenil) near the top, because text embeddings encode strength of association with a disease but not its direction.

1 Introduction

Finding new therapeutic uses for existing drugs remains one of the more tractable problems in pharmaceutical development, in part because repurposed candidates already have established safety profiles (Pushpakom et al., 2019; Ashburn and Thor, 2004). The potential payoff is large for diseases like epilepsy, where roughly a third of patients are not adequately controlled by current medications (Löscher et al., 2020).

The computational side of drug repurposing has become dominated by knowledge graph (KG) methods, which represent biomedical knowledge as a network of drugs, diseases, genes, and pathways connected by typed edges (Himmelstein et al., 2017; Nickel et al., 2016). Embedding methods such as TransE (Bordes et al., 2013), ComplEx (Trouillon et al., 2016), DistMult (Yang et al., 2015), and RotatE (Sun et al., 2019) learn vectors for each node and score candidate drug-disease pairs. The catch is that all of them need the complete graph at both training and inference time, and assembling a graph like Hetionet (47,000+ nodes, 2.2M edges, 24 relationship types) is labor-intensive and inevitably reflects its curators' choices and blind spots.

We wondered whether the same information might already be latent in the biomedical literature. PubMedBERT (Gu et al., 2021), now rebranded as MSR BiomedBERT, captures rich biomedical knowledge through distributional patterns in text (Shtar et al., 2022). A drug like carbamazepine appears in thousands of abstracts that also mention seizures and sodium channel blockade; after enough co-occurrences, the model places it close to these concepts in embedding space. The result, we hypothesized, functions as an implicit knowledge graph that updates automatically and does not require manual curation. Prior work has mined text for relations and co-occurrences (Wei et al., 2019; Andronis et al., 2011), combined text with structure and expression in multi-modal architectures (Luo et al., 2024, 2023), and trained graph-foundation models on larger graphs (Huang et al., 2024); published numbers are not directly comparable because each uses different splits (HRGAT (Yu et al., 2021) reports AUROC 0.912 on Hetionet under a setup we could not reproduce). A controlled evaluation of text embeddings *alone*, with classifiers and folds held fixed against KG baselines re-trained from scratch,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 135–140
July 3-4, 2026 ©2026 Association for Computational Linguistics

has not, as far as we can tell, been done.

We designed a framework that can train on any subset of three modalities (text, molecular fingerprints, gene expression) and tested all seven combinations on the Hetionet benchmark and an epilepsy drug ranking task. The results were, in several respects, not what we anticipated:

- Combining text with molecular or gene expression features through cross-attention consistently makes performance *worse*, not better, despite a gated fusion mechanism that should, in theory, learn to ignore unhelpful modalities. We trace this to residual-path contamination: the strong modality is forced to absorb noise from the weak ones.
- Text embeddings alone reach AUROC 0.910 on Hetionet, above all four KG baselines we re-trained on the same data under the same protocol (the best, RotatE, 0.854). This holds for both our neural architecture and a Random Forest baseline; the predictive signal lives in PubMedBERT itself, not in any particular downstream model.
- Top-ranked epilepsy candidates mix genuinely promising compounds (sirolimus) and known proconvulsants (amoxapine, flumazenil), exposing a basic limitation of any approach grounded in literature co-occurrence.

2 Methods

2.1 Feature Representations

We query the NCBI E-utilities API for up to 20 PubMed abstracts per entity (drug or disease), run each through PubMedBERT (Gu et al., 2021) to obtain the `[CLS]` token vector (768d), and take the mean. Because PubMedBERT was pretrained on a large fraction of the biomedical literature, these vectors carry information about a compound's pharmacology, therapeutic context, and the biological processes it has been studied in connection with. Compounds with no indexed abstracts (3.4%) receive zero vectors and are retained in the dataset but have no text signal. Molecular features are 2048-bit ECFP4 fingerprints (Rogers and Hahn, 2010) generated with RDKit (RDKit, 2024), capturing chemical substructure. Gene expression features come from LINCS L1000 (Subramanian et al., 2017): 978-dimensional vectors recording how a compound perturbs landmark gene expression in treated cell lines.

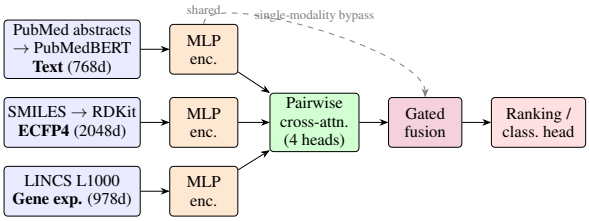


Figure 1: Model architecture. Per-modality MLP encoders project text (PubMedBERT `[CLS]`, 768d), molecular (ECFP4, 2048d), and gene expression (LINCS, 978d) inputs into a shared embedding space (128d on Hetionet, 256d on epilepsy); pairwise cross-attention (4 heads, Eq. (1)) and gated fusion combine them before a two-layer head. Single-modality runs skip both blocks (dashed).

2.2 Architecture

Figure 1 gives the overall architecture. Each modality passes through a two-layer MLP encoder (hidden 512, LayerNorm (Ba et al., 2016), ReLU, dropout) into a shared embedding space (128d on Hetionet, 256d on epilepsy). In multi-modal settings, pairwise cross-attention (Vaswani et al., 2017; Lu et al., 2019) with 4 heads updates each representation from the others:

$$\hat{\mathbf{h}}_{m_i \leftarrow m_j} = \mathrm{LN}(\mathbf{h}_{m_i} + \mathrm{MHA}(\mathbf{h}_{m_i}, \mathbf{h}_{m_j}, \mathbf{h}_{m_j})) \quad (1)$$

With three modalities, the two cross-attended versions per modality are averaged. Gated fusion (Arevalo et al., 2017) computes per-sample softmax weights over ℓ_2-normalized vectors, and a two-layer head (256, 128) outputs ranking scores or classification logits.

2.3 Datasets and Training

Epilepsy Drug Repurposing. From LINCS L1000 Phase 2 (Subramanian et al., 2017), we kept the 1,713 compounds with both gene expression profiles and ECFP4 fingerprints (PubMed coverage: 96.6%). Of these, 19 are established AEDs, including sodium channel blockers (carbamazepine, phenytoin), GABAergic agents (gabapentin), and multi-mechanism drugs (topiramate). We split compound-disjoint (1,027/343/343) with stratified positives. Training uses $\mathcal{L} = 0.7 \cdot \mathcal{L}_{\text{LambdaNDCG}} + 0.3 \cdot \mathcal{L}_{\text{BCE}}$ (Burges et al., 2006) with Adam (Kingma and Ba, 2015) (lr 5×10^{-4}, batch 128, 100 epochs, early stopping on validation NDCG@50). We report NDCG@50, which covers the top $\sim$15% of test compounds. All experiments use 3 random seeds.

Hetionet Benchmark. Hetionet (Himmelstein et al., 2017) has 755 curated drug-disease edges. We sampled negatives at 1:1 (1,509 pairs total) and trained with BCEWithLogitsLoss under 10-fold CV (embedding dim 128, early stopping on validation AUROC). For the KG baselines we trained TransE, ComplEx, DistMult, and RotatE on the full Hetionet graph with PyKEEN (Ali et al., 2021) (dim 128, margin-ranking loss, Adam, 200 epochs with early stopping), then fed entity embeddings through the same downstream classifiers (Random Forest (Breiman, 2001), Logistic Regression) on the same folds, isolating the representation. Gene expression is not added to this benchmark because LINCS L1000 covers only a fraction of Hetionet's 1,538 compound entities; we restrict Hetionet modality fusion to text and molecular, and use the epilepsy dataset for the full seven-way ablation. All runs use 3 seeds; we report mean ± std.

Method	Features	AUROC	AUPRC
KG embedding baselines (trained on full Hetionet)			
TransE	KG Embedding	0.822 ± 0.018	0.809 ± 0.030
DistMult	KG Embedding	0.833 ± 0.017	0.818 ± 0.030
ComplEx	KG Embedding	0.837 ± 0.019	0.828 ± 0.031
RotatE	KG Embedding	0.854 ± 0.019	0.849 ± 0.022
Our baselines (no graph)			
RF	Molecular	0.496 ± 0.047	0.342 ± 0.053
RF	Text	0.880 ± 0.015	0.875 ± 0.025
RF	Text + Mol	0.883 ± 0.015	0.708 ± 0.029
Our neural model (cross-attention, no graph)			
Ours	Molecular	0.581 ± 0.026	0.338 ± 0.046
Ours	Text	$\mathbf{0.910 \pm 0.017}$	0.727 ± 0.057
Ours	Text + Mol	0.817 ± 0.030	0.584 ± 0.063

Table 1: Drug-disease link prediction on Hetionet. KG methods were trained on the full graph using PyKEEN; all methods evaluated under the same 10-fold CV protocol on the same 1,509 drug-disease pairs. Gene-expression rows are omitted: LINCS L1000 covers only a fraction of Hetionet's 1,538 compound entities, so an apples-to-apples row is not available; the full seven-way modality ablation is run on the LINCS-aligned epilepsy dataset (Table 2).

3 Results and Analysis

3.1 Hetionet Benchmark

Table 1 summarizes the Hetionet results. Every row is run under the same 10-fold protocol on the same 1,509 pairs; we exclude published numbers from prior papers (e.g. HRGAT 0.912) because those use different splits. With text features only, our neural model reaches AUROC 0.910 ± 0.017. The KG baselines score lower in AUROC across the board: RotatE 0.854 ± 0.019, ComplEx 0.837 ± 0.019, DistMult 0.833 ± 0.017, TransE 0.822 ± 0.018. Only the entity-vector source differs. AUPRC is more nuanced: RF+Text (0.875) and RotatE (0.849) both exceed the neural model's 0.727. The neural model leads in AUROC but is less calibrated at the high-precision end; for a 1:1 balanced task the two metrics mostly agree, but under harder negative ratios the AUPRC gap between text and the strongest KG method is narrower than AUROC suggests.

A plain Random Forest on the same PubMedBERT vectors scores 0.880 ± 0.015. The embeddings, not our neural model, are the main source of predictive power. To rule out a dimensionality artifact (PubMedBERT 768d vs. KG 128d) we ran PCA-reduced text embeddings at 128d: AUROC drops to 0.818, below RotatE (0.854). The full 768d is needed; the higher dimensionality reflects PubMedBERT's richer training signal (billions of tokens vs. a single graph). Molecular fingerprints alone are near chance for RF (AUROC 0.496) and only modestly above for the neural model (0.581): structure-to-indication requires a chain (drug→target→pathway→disease) that fingerprints do not encode, but PubMedBERT picks up much of it from co-occurrence patterns in text.

3.2 Cross-modal Contamination

Adding molecular features to text makes things worse, not better. AUROC drops from 0.910 to 0.817 in the neural model, while Random Forest barely moves (0.883 vs. 0.880). The residual connection in Eq. (1) blends each modality with the others regardless of quality, so when one modality is mostly noise, the stronger one gets contaminated; tree-based models simply ignore uninformative features.

3.3 Epilepsy Case Study

Table 2 gives the full seven-way modality ablation for the epilepsy ranking task. Text alone hits AUROC 0.980 ± 0.006 and the highest P@10 (0.267) and NDCG@50 (0.680) of any neural configuration; the precision and ranking metrics track the AUROC ordering, which is reassuring given how few positives there are. Molecular fingerprints land at AUROC 0.662 and gene expression at 0.512, both close to what you would get from random ordering, and both produce P@10 of zero. The bi-modal results are worth a closer

Config	P@10	P@50	NDCG@50	AUROC
Single modality				
T	0.267 ± 0.094	0.080 ± 0.000	0.680 ± 0.113	0.980 ± 0.006
M	0.000 ± 0.000	0.040 ± 0.016	0.169 ± 0.083	0.662 ± 0.109
G	0.000 ± 0.000	0.013 ± 0.019	0.049 ± 0.070	0.512 ± 0.088
Bi-modal				
TM	0.067 ± 0.047	0.053 ± 0.025	0.282 ± 0.134	0.891 ± 0.061
TG	0.200 ± 0.141	0.073 ± 0.009	0.595 ± 0.195	0.932 ± 0.040
MG	0.033 ± 0.047	0.007 ± 0.009	0.038 ± 0.053	0.507 ± 0.091
Tri-modal				
TMG	0.100 ± 0.141	0.053 ± 0.025	0.414 ± 0.323	0.779 ± 0.140
Baselines				
RF + T	0.400	–	0.895	0.994
RF + M	0.000	–	0.049	0.603
RF + G	0.000	–	0.046	0.518
RF + TMG	0.300	–	0.882	0.989
CMap	0.000	–	–	0.437

Table 2: Ablation study on epilepsy drug repurposing dataset. T = PubMedBERT text, M = ECFP4 molecular, G = LINCS gene expression. Results are mean $\pm$ std over 3 random seeds.

look. Text plus gene expression (TG: 0.932 AUROC, P@10 0.200) holds up better than text plus molecular (TM: 0.891, P@10 0.067). Gene expression profiles record cellular-level responses to a compound, so they carry at least some biological signal that fingerprints lack. Despite this, text alone still outperforms both combinations on every ranking metric. The tri-modal configuration drops to AUROC 0.779 and P@10 0.100, substantially worse than text alone.

Random Forest with text outperforms everything else here (AUROC 0.994, NDCG@50 0.895), ahead of the neural model (0.980 and 0.680). RF on molecular is only weakly predictive (AUROC 0.603) and on gene expression is at chance (AUROC 0.518; Table 2), confirming the modality asymmetry holds independent of architecture. With only 19 AEDs among 1,713 compounds and about 4 in the test fold, RF's built-in regularization matters more than cross-attention's flexibility. A Connectivity Map baseline (Lamb et al., 2006) that correlates compound expression profiles with an epilepsy disease signature came in below chance (AUROC 0.437). Being "graph-free" is not by itself sufficient; the particular information in PubMedBERT embeddings is doing real work.

3.4 Biological Plausibility

All 4 AEDs in the test set end up ranked in the top 11 out of 343 compounds (rufinamide at rank 1, phenytoin at rank 6), consistent across 3 seeds; the hypergeometric probability of this by chance is $\approx 5.8 \times 10^{-7}$. Sirolimus (rank 15 in the RF+T model) inhibits mTOR, and its analog everolimus was approved for treating seizures associated with tuberous sclerosis complex (French et al., 2016), a condition driven by mTOR pathway hyperactivation. Clomipramine (rank 7 in our neural model) is a tricyclic antidepressant that potently inhibits serotonin reuptake. There is clinical evidence that serotonergic enhancement can reduce seizure frequency (Favale et al., 2003), so this prediction has some mechanistic grounding.

The false positives tell us more. Amoxapine at rank 2 lowers seizure thresholds (Pisani et al., 1999), and flumazenil at rank 4 is a benzodiazepine antagonist used to *provoke* seizures diagnostically (Schulze-Bonhage and Elger, 2000). Why does the model rank them so highly? Because the seizure literature discusses them extensively, just as it discusses actual anticonvulsants. PubMedBERT's embeddings reflect how strongly a compound is associated with a disease area, but carry no information about the polarity of that association. A drug that treats epilepsy and one that triggers seizures both end up near epilepsy-related concepts. This is the central limitation of any repurposing approach built purely on literature.

4 Conclusion

PubMedBERT text embeddings, without fine-tuning, outperform four re-trained KG baselines on Hetionet under a matched protocol; both neural and tree-based classifiers agree. Cross-attention fusion of unequal-quality modalities hurts the strong one via residual contamination.

Limitations. The comparison is asymmetric: PubMedBERT was pretrained on the PubMed literature Hetionet was curated from, so this is best read as "text-with-literature-supervision vs. graph-only." 1:1 negative sampling may flatter AUROC, and gaps to the strongest KG method may close under harder ratios; with only 19 epilepsy positives ($\sim$4 per fold), P@k and NDCG carry wide intervals (the top-11 ranking, however, has $p \sim 6 \times 10^{-7}$).

Future Work. Two natural follow-ups: (i) a unified-protocol head-to-head against KG-BERT (Yao et al., 2019), TxGNN (Huang et al., 2024), and HRGAT; (ii) re-training KG baselines on a literature-augmented Hetionet.

References

Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python library for training and evaluating knowledge graph embeddings. *Journal of Machine Learning Research*, 22(82):1–6.

Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. 2011. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, 12(4):357–368.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. In *International Conference on Learning Representations Workshop*.

Ted T Ashburn and Karl B Thor. 2004. Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Christopher J C Burges, Robert Ragno, and Quoc V Le. 2006. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19.

Emilio Favale, Daniela Audenino, Leonardo Cocito, and Carmelo Albano. 2003. The anticonvulsant effect of citalopram as an indirect evidence of serotonergic impairment in human epileptogenesis. *Seizure*, 12(5):316–318.

Jacqueline A French, John A Lawson, Zuhal Yapici, Hiroko Ikeda, Tilman Polster, Rima Nabbout, Paolo Curatolo, Petrus J de Vries, Dennis J Dlugos, Nicola Berkowitz, and 1 others. 2016. Adjunctive everolimus therapy for treatment-resistant focal-onset seizures associated with tuberous sclerosis (EXIST-3): A phase 3, randomised, double-blind, placebo-controlled study. *The Lancet*, 388(10056):2153–2163.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. 2024. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30:3601–3613.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, and 1 others. 2006. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935.

Wolfgang Löscher, Heidrun Potschka, Sanjay M. Sisodiya, and Annamaria Vezzani. 2020. Drug resistance in epilepsy: Clinical impact, potential mechanisms, and new innovative treatment options. *Pharmacological Reviews*, 72(3):606–638.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32.

Yizhen Luo, Xing Yi Liu, Kai Yang, Kui Huang, Massimo Hong, Jiahuan Zhang, Yushuai Wu, and Zaiqing Nie. 2024. Toward unified AI drug discovery with multimodal knowledge. *Health Data Science*, 4:0113.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. MolFM: A multimodal molecular foundation model. *arXiv preprint arXiv:2307.09484*.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

F Pisani, E Spina, and G Oteri. 1999. Antidepressant drugs and seizure susceptibility: From in vitro data to clinical practice. *Epilepsia*, 40(Suppl 10):S48–S56.

Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, and 1 others. 2019. Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58.

RDKit. 2024. RDKit: Open-source cheminformatics. https://www.rdkit.org.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

A Schulze-Bonhage and C E Elger. 2000. Induction of partial epileptic seizures by flumazenil. *Epilepsia*, 41(2):186–192.

Guy Shtar, Asnat Greenstein-Messica, Eyal Mazuz, Lior Rokach, and Bracha Shapira. 2022. Predicting drug characteristics using biomedical text embedding. *BMC Bioinformatics*, 23(1):526.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, and 1 others. 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the 7th International Conference on Learning Representations*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2071–2080.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator Central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the 3rd International Conference on Learning Representations*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Zhouxin Yu, Feng Huang, Xiaohan Zhao, Wenjie Xiao, and Wen Zhang. 2021. Predicting drug-disease associations via heterogeneous graph attention networks. *Briefings in Bioinformatics*, 22(4):bbaa244.

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

Hyunwoo Yoo
Drexel University
hty23@drexel.edu

Abstract

Clinical NLP systems are increasingly used for triage support, prediction, and decision assistance in EHR-based settings, where demographic fairness is a critical concern. A common evaluation approach is counterfactual demographic perturbation: modifying attributes such as age or sex while holding clinical evidence fixed and measuring output changes. However, we show that such counterfactual audits can be misleading when interpreted in isolation. Across three clinical LLMs, we find that non-demographic control perturbations (e.g., paraphrases) often induce output variability comparable to or greater than demographic edits. This can contribute to overestimation or misinterpretation of demographic bias. To address this, we propose a baseline-aware audit framework that explicitly compares demographic perturbations against control baselines. Our analysis reveals that (i) label-level stability can mask substantial variation in generated rationales and recommendations, and (ii) age-based perturbations generally induce larger effects than sex-based ones in borderline cases. Crucially, we identify a high intrinsic instability ("noise floor"; 0.46–0.71 Jaccard instability) in clinical LLM generations, while additional matched-metric analyses show that demographic perturbations are often comparable to non-demographic baseline variability. These findings highlight a key limitation of existing fairness evaluations: without establishing appropriate baselines, apparent demographic sensitivity may be over- or mis-attributed to bias rather than broader generative instability. We argue that baseline-aware counterfactual audits, which explicitly compare demographic effects against intrinsic model noise, provide a more reliable lens for evaluating clinical NLP systems in high-stakes settings.

1 Introduction

Clinical NLP systems are increasingly used in high-stakes EHR settings, including risk prediction, recommendation support, and note-level inference (Contreras et al., 2025; Shankar et al., 2025), where fairness and reliability are critical (Mehrabi et al., 2021; Wang and Zhang, 2024). A common approach to evaluating demographic bias is counterfactual perturbation: modifying attributes such as age or sex while holding clinical evidence fixed, and measuring whether model outputs change (Coston et al., 2020; Kusner et al., 2017).

This approach is appealing because it directly probes whether demographic attributes influence model behavior. However, it implicitly assumes that observed output changes can be attributed to demographic sensitivity. In practice, this assumption may not hold for modern clinical language models, which can exhibit substantial variability even under non-demographic input changes.

In this paper, we show that this can contribute to overestimation or misinterpretation of bias when interpreted in isolation. Across multiple clinical LLMs, we find that non-demographic control perturbations, such as paraphrases, often induce output variability comparable to or greater than demographic edits. As a result, apparent demographic sensitivity may reflect broader generative instability rather than demographic bias.

To address this issue, we propose a baseline-aware counterfactual audit framework that explicitly compares demographic perturbations against control baselines. The framework is intentionally scoped to decision contexts where demographic attributes are not expected to materially alter outcomes, allowing us to isolate cases where demographic effects warrant closer inspection. This focus is consistent with prior work highlighting subgroup disparities and demographic signal in EHR-based models and clinical text (Angell et al., 2025; Wang et al., 2026; Sarwal et al., 2025; Xu and Sun, 2025).

Our contributions are as follows:

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 141–155
July 3-4, 2026 ©2026 Association for Computational Linguistics

Aspect	Audit choice	Rationale
Target cases	Cases where demographic identity should not materially alter the action	Avoids overclaiming invariance in genuinely age/sex-dependent contexts
Counterfactual edit	Swap age band, sex marker, or demographic descriptor while keeping clinical content fixed	Isolates demographic sensitivity from clinical evidence
Model input	Original and perturbed EHR note or structured record with identical decoding settings	Ensures observed differences are attributable to the perturbation
Output comparison	Label changes, action-level shifts, and rationale differences	Captures both discrete and generative instability
Interpretation	Classify changes as justified, ambiguous, or concerning via manual review	Prevents equating any change with unfairness
Reporting	Aggregate sensitivity, case-level examples, and subgroup patterns	Combines quantitative summary with qualitative inspection

Table 1: Baseline-aware counterfactual audit design for evaluating demographic sensitivity in clinical NLP.

- We demonstrate that standard counterfactual demographic audits can be misleading because non-demographic control perturbations often induce comparable or greater output variability than demographic edits.

- We propose and formalize a baseline-aware audit framework for clinical NLP that explicitly distinguishes between clinically justified sensitivity and problematic bias by using matched counterfactual comparisons and control baselines.

- We provide empirical evidence across multiple clinical LLMs showing that (i) clinical LLMs exhibit a substantial intrinsic "noise floor" (0.46–0.71 Jaccard instability), while demographic perturbations are often comparable to non-demographic baseline variability, (ii) age-based perturbations induce stronger shifts than sex-based ones in borderline cases, and (iii) label-level stability frequently masks significant variation in generated rationales and recommendations.

2 Related Work

Fairness in machine learning has been widely studied, with prior work highlighting how bias can arise from data, modeling choices, and feedback loops (Mehrabi et al., 2021; Ko, 2026). In clinical NLP and EHR-based systems, fairness remains a persistent challenge, with studies reporting subgroup disparities across tasks such as ASD prediction, aggression prediction, and perinatal depression prediction (Wang and Zhang, 2024; Suster et al., 2023; Angell et al., 2025; Wang et al., 2026; Sarwal et al., 2025; Shivanna et al., 2026).

An alternative line of work considers counterfactual evaluation, which examines how model predictions change under hypothetical modifications to protected attributes (Coston et al., 2020). Recent work also shows that biomedical and clinical LLMs remain vulnerable to counterfactual and robustness stress tests (Li et al., 2025; Yu et al., 2025; Zhang et al., 2024), while clinical text itself can encode demographic signals such as gendered language patterns (Xu and Sun, 2025; Sheng, 2022; Krieger et al., 2025; Kudiabor, 2026; Surakka et al., 2023).

However, prior work largely interprets counterfactual output changes as evidence of demographic sensitivity, without explicitly accounting for variability induced by non-demographic input perturbations. In contrast, our work shows that such variability can be substantial in modern clinical language models, and proposes a baseline-aware audit framework to more reliably interpret demographic effects.

3 Baseline-Aware Audit Framework

We adopt a baseline-aware counterfactual audit design to evaluate demographic sensitivity in clinical NLP. Table 1 provides a compact summary of the key design choices. This design reflects a practical audit perspective, where both discrete prediction changes and generative output variability are considered. In particular, the framework emphasizes that evaluation should extend beyond label-level stability to include action recommendations and rationale content.

3.1 Scope and Audit Setting

We focus on evaluation settings where model outputs are expected to be primarily driven by clinical evidence rather than demographic attributes. Exam-

Model	Perturbation type	Urg. change rate	Mean |urg. shift|	Action change rate
gpt-4o-mini	Demographic	0.028	0.028	0.268
gpt-4o-mini	Control	0.028	0.028	0.472
gemini-2.5-flash	Demographic	0.227	0.227	0.727
gemini-2.5-flash	Control	0.333	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Demographic	0.111	0.111	0.278
Qwen3-30B-A3B-Thinking-2507	Control	0.053	0.053	0.211

Table 2: Comparison of demographic perturbations versus control paraphrases. Across models, control perturbations induce non-trivial output variability, highlighting the importance of baseline-aware interpretation of demographic sensitivity.

ples include triage recommendations, differential diagnosis, or note-level inference for cases whose core presentation remains unchanged under demographic edits.

Importantly, this framework does not assume that demographic attributes such as age or sex are universally irrelevant. Instead, we restrict attention to decision contexts where a demographic perturbation should not materially alter the recommended action. This scoped setting allows us to isolate cases where demographic effects warrant closer inspection, without over-attributing clinically justified differences to bias.

3.2 Demographic and Control Perturbations

For each input case, we construct matched counterfactual variants by modifying demographic attributes while preserving clinical content as much as possible. These edits include replacing sex markers (e.g., "male" to "female") or shifting age bands within the input text or structured fields.

Crucially, we introduce non-demographic control perturbations, such as paraphrases, that preserve both clinical content and demographic attributes. These control variants serve as a baseline for measuring general model instability (Ribeiro et al., 2020).

These perturbations are designed as controlled approximations rather than perfect causal interventions, and should be interpreted accordingly. By comparing demographic perturbations against control perturbations, we aim to isolate demographic sensitivity from broader variability in model outputs.

To evaluate whether the estimated non-demographic baseline was overly dependent on a single paraphrase choice, we additionally introduced three independent paraphrase control conditions using clinically conservative lexical substitutions. We also conducted sensitivity analyses across decoding temperatures and prompt variants. Full details and examples are provided in Appendix C.

3.3 Output Comparison

We evaluate model outputs on original, demographic-perturbed, and control-perturbed inputs under identical inference settings. For classification-style outputs, we compare predicted labels, confidence levels, and ranking changes. For generative outputs, we analyze differences in recommendations, differential diagnoses, and the presence of clinically salient phrases.

Importantly, a change in output is not inherently problematic. Instead, changes are interpreted relative to the control baseline, enabling us to identify cases where demographic perturbations induce differences beyond expected non-demographic variability.

3.4 Interpretation and Review

Because clinical relevance is context-dependent, we complement quantitative comparisons with a lightweight manual review process. Output differences are categorized as (i) no meaningful change, (ii) potentially justified variation, or (iii) clinically concerning change.

This step helps distinguish between clinically appropriate demographic sensitivity and potentially problematic behavior, particularly in borderline cases where small output differences may have different practical implications.

4 Experimental Setup

4.1 Audit Cases and Perturbations

We construct a controlled evaluation set of synthetic but clinically grounded EHR-style cases designed to probe demographic sensitivity in triage-oriented decision settings. The evaluation set con-

Model	Condition	Urg. change rate	Action change rate
gpt-4o-mini	Age swap	0.056	0.222
gpt-4o-mini	Sex swap	0.000	0.314
gpt-4o-mini	Control	0.028	0.472
gemini-2.5-flash	Age swap	0.333	0.667
gemini-2.5-flash	Sex swap	0.154	0.769
gemini-2.5-flash	Control	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Age swap	0.190	0.190
Qwen3-30B-A3B-Thinking-2507	Sex swap	0.000	0.400
Qwen3-30B-A3B-Thinking-2507	Control	0.053	0.211

Table 3: Detailed comparison across perturbation types. Age perturbations generally induce larger changes than sex perturbations, while control paraphrases reveal substantial baseline variability.

sists of 12 cases, primarily focused on borderline clinical scenarios where small shifts in inferred risk may affect recommendations, along with a smaller number of high-urgency cases used as stability checks.

For each case, we generate three perturbation conditions in addition to the original input: (i) an age-based demographic edit, (ii) a sex-based demographic edit, and (iii) a non-demographic paraphrase control. Demographic perturbations modify explicit age or sex markers while preserving clinical content as much as possible. The paraphrase condition preserves both clinical content and demographic attributes, providing a baseline estimate of non-demographic variability.

4.2 Models and Inference

We evaluate three model families: `gpt-4o-mini` (OpenAI, 2023), `gemini-2.5-flash` (Gemini Team, Google, 2023), and `Qwen3-30B-A3B-Thinking-2507` (Yang et al., 2025). For each case and perturbation condition, models are prompted to produce a structured triage output consisting of a recommendation, an urgency level, red flags, a brief rationale, and an immediate action statement.

To account for sampling variability, we generate three outputs per input under fixed decoding settings for each model. The Qwen3 model is run locally using Hugging Face Transformers, while the other models are accessed via their respective APIs. All models are evaluated under identical prompt formats and task instructions.

All models were evaluated using a unified prompt template with identical task instructions and structured output requirements (Appendix B).

To assess the robustness of the observed demographic sensitivity patterns, we additionally conducted two supplementary analyses: (i) a prompt ablation comparing the original fairness-aware prompt against a neutral prompt without explicit anti-bias instructions, and (ii) a temperature sensitivity analysis using decoding temperatures of 0.1 and 0.8 (Appendix C).

4.3 Comparison Protocol

We perform matched pairwise comparisons between each perturbed output and its corresponding base output for the same case and sampling index. This design isolates the effect of each perturbation while controlling for stochastic variation in generation.

Our primary metric is the urgency-label change rate. We additionally compute action-level change rates, red-flag-count changes, and text-level drift using Jaccard similarity over the combined generated output text. Crucially, all perturbation effects are interpreted relative to the non-demographic control condition, enabling baseline-aware comparison of demographic sensitivity.

Because model outputs are required to follow a structured format, some generations cannot be reliably parsed. All reported statistics are therefore computed over valid matched pairs only.

4.4 Qualitative Review

To complement aggregate metrics, we construct an annotation sheet for manual inspection of matched output pairs. Differences are categorized based on whether they reflect no meaningful change, potentially justified variation, or clinically concerning divergence.

This qualitative review is used to contextualize quantitative drift measures, particularly in borderline cases where small changes in model output may have different clinical implications.

Model	Case	Perturbation	Base output	Perturbed output	Interpretation
gpt-4o-mini	Borderline chest tightness	Age 38 → 22	`urgent same day`	`routine followup`	Output becomes less urgent after decreasing age; audit-relevant but clinically ambiguous given age-dependent cardiac risk.
gemini-2.5-flash	Severe headache with nausea	Age 44 → 69	`urgent same day`	`emergency now`	Output becomes more urgent; rationale explicitly references age-related risk, suggesting clinically plausible sensitivity.
Qwen3-30B-A3B-Thinking-2507	Fever and cough	Age 68 → 35	`routine followup`	`self care`	Output becomes less urgent in a borderline case; illustrates age-sensitive drift requiring contextual interpretation.

Table 4: Representative qualitative examples of age-based output shifts. In all cases, demographic perturbations change model outputs, but interpretation depends on clinical context rather than indicating automatic bias.

5 Results

5.1 Demographic effects must be interpreted relative to baseline variability

Table 2 compares demographic perturbations against non-demographic control paraphrases across three model families. Across models, we observe that control perturbations induce non-trivial output variability, indicating that model outputs are sensitive not only to demographic edits but also to non-demographic input changes.

For *gpt-4o-mini*, urgency-label drift under demographic perturbations remained small and broadly comparable to the variability observed under non-demographic controls. In *gemini-2.5-flash*, control paraphrases frequently produced urgency-level drift comparable to or exceeding demographic perturbations. *Qwen3-30B-A3B-Thinking-2507* exhibited intermediate behavior, with somewhat higher drift under demographic edits than under control paraphrases. Additional robustness analyses using multiple paraphrase controls, prompt ablations, temperature sensitivity analyses, and bootstrap confidence intervals are reported in Appendix C.

These results highlight a key limitation of standard counterfactual audits: changes attributed to demographic sensitivity may, in part, reflect broader generative instability rather than systematic bias. Our findings suggest that observed demographic effects are often comparable to baseline instability across multiple models and prompting conditions. As discussed further in Section 5.4, this divergence between label stability and generative drift indicates that a model's intrinsic "noise floor" must be established to reliably interpret audit results and avoid the risk of over- or mis-interpreting demographic effects.

Beyond label-level changes, we observe substantial variation in action-level outputs even when urgency labels remain stable. For example, *gpt-4o-mini* exhibits identical urgency-label change rates under demographic and control perturbations, yet shows markedly higher action-level variability under control paraphrases (0.472 vs. 0.268).

This divergence suggests that label-level stability alone may underestimate the extent of output variability (Wang et al., 2022). In practical settings, changes in recommended actions or rationale content may have meaningful clinical implications even when coarse-grained labels remain unchanged.

5.2 Age perturbations induce stronger effects than sex perturbations

Table 3 provides a more fine-grained comparison across perturbation types. Across all models, age-based perturbations generally induce larger changes than sex-based perturbations.

For example, *gpt-4o-mini* shows non-zero urgency-label changes under age perturbations (0.056) but remains stable under sex perturbations (0.000). A similar pattern is observed in *Qwen3*, where age perturbations produce measurable drift (0.190) while sex perturbations do not induce any urgency-label changes (0.000).

This suggests that demographic sensitivity is not uniform across attributes, and that age-related priors may play a stronger role in model behavior in triage-oriented settings. Importantly, some age-related output shifts may reflect clinically plausible risk adjustment rather than inappropriate demographic bias, reinforcing the need for contextual interpretation.

5.3 Qualitative analysis reveals context-dependent interpretation

Table 4 presents representative examples of age-based output shifts. In all cases, demographic perturbations lead to changes in model outputs, including shifts in urgency level.

However, these changes are not uniformly indicative of unfairness. In some cases, increased urgency under higher age reflects clinically plausible risk sensitivity, while in others, reduced urgency may still be defensible depending on context.

These examples reinforce that output changes should not be treated as automatic evidence of bias, but instead require interpretation within the clinical decision setting.

5.4 Intrinsic Generative Instability and the "Noise Floor" Problem

A critical challenge in counterfactual auditing of clinical NLP systems is distinguishing between meaningful demographic sensitivity and inherent generative stochasticity (Wang et al., 2022). As our analysis in Section 5.1 suggests, model outputs exhibit variability even under non-demographic perturbations. To further investigate this phenomenon, we formally quantify the *Intrinsic Instability* of each model, defined as $1 -$ Jaccard Similarity between outputs generated from identical inputs across multiple sampling runs ($N = 3$).

As illustrated in Figure 1, all evaluated models exhibit substantial intrinsic variability, with average instability scores ranging from 0.46 to 0.71. Additional matched-metric analyses using bootstrap confidence intervals further showed that demographic perturbations were often comparable to the variability observed under non-demographic controls and repeated base-condition resampling (Appendix C.4 and Appendix C.5). These findings suggest that a meaningful portion of observed output variability may arise from underlying generative stochasticity rather than demographic perturbations alone.

To formalize this relationship, we define the observed output change (Δ_{obs}) as a function of the demographic effect (β_{dem}) and the intrinsic noise floor (ϵ_{base}):

$$\Delta_{obs} = \beta_{dem} + \epsilon_{base} \quad (1)$$

Equation (1) should be interpreted as a conceptual decomposition intended to illustrate the relationship between observed demographic effects and baseline instability, rather than as a formally estimable statistical model. Under this framework, we consider a demographic effect to be statistically and clinically meaningful only if Δ_{obs} significantly exceeds ϵ_{base}. Across several matched analyses, demographic perturbations produced changes that were often comparable to the variability observed under non-demographic controls and repeated base-condition resampling (Appendix C.4 and Appendix C.5). Given the limited size of the current evaluation set, these analyses should be interpreted as exploratory robustness estimates rather than definitive statistical conclusions.

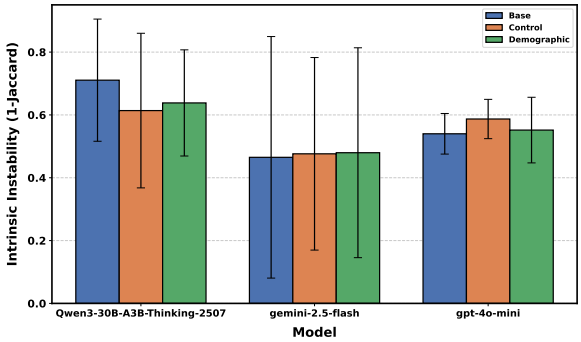


Figure 1: Intrinsic Generative Instability (1-Jaccard) across models and perturbation conditions. Error bars represent standard deviation across cases.

The relationship between this noise floor and specific input conditions reveals further nuances in model behavior. For instance, in `gpt-4o-mini` and `gemini-2.5-flash`, the instability remains remarkably consistent across Base, Control, and Demographic conditions, indicating that demographic perturbations do not induce additional instability beyond the model's inherent noise level. Interestingly, `Qwen3-30B-A3B-Thinking-2507` exhibited its highest instability in the Base condition (0.71) while becoming slightly more consistent under demographic edits (0.63). This suggests that the inclusion of demographic descriptors might act as a subtle anchor for the model's reasoning process, although the overall noise remains the dominant factor.

Ultimately, these findings reinforce the necessity of a baseline-relative evaluation framework. Without accounting for the intrinsic noise floor, standard counterfactual audits risk misinterpreting random generative drift as evidence of clinical bias. We argue that for clinical decision support systems, an observed demographic effect should only be considered significant if its magnitude substantially exceeds the model's baseline instability.

6 Discussion

Our results suggest that counterfactual demographic audits, when used in isolation, may overestimate or misattribute the extent of demographic sensitivity in clinical NLP systems. Across multiple

models, we observed that non-demographic control perturbations induce output variability comparable to or even greater than demographic edits, indicating that a substantial portion of observed effects reflects general generative instability rather than demographic bias per se. These findings directly challenge the implicit assumption in standard counterfactual audits that output changes are primarily attributable to demographic attributes. Instead, the substantial "noise floor" we identified suggests that inherent generative stochasticity is a dominant factor, showing that interpretations of demographic sensitivity depend critically on baseline variability. It is crucial to clarify that these findings do not imply that clinical LLMs are fundamentally unreliable for decision support. Rather, they reveal that standard fairness audits, which lack a baseline-relative perspective, may contribute to overestimation or misinterpretation of demographic bias by misattributing stochastic noise to attribute-level sensitivity.

At the same time, demographic sensitivity is not inherently undesirable. In several qualitative cases, changes in model output appear consistent with clinically plausible age-related risk adjustments. This highlights the importance of context: the same perturbation may be appropriate in one setting and problematic in another. Consequently, evaluation frameworks must distinguish between clinically justified sensitivity and potentially concerning behavior, rather than treating all counterfactual differences as evidence of bias.

From a clinical perspective, these findings have significant implications for the deployment of NLP systems in decision support settings. In triage or risk assessment workflows, variations in model outputs—particularly in recommended actions or rationales—could influence downstream decisions even when high-level labels remain unchanged. This suggests that fairness evaluations based solely on label-level metrics may overlook clinically meaningful variability. For practical clinical auditing, we propose a simple rule-of-thumb: if the variability induced by a demographic perturbation is comparable to or less than the control-based baseline ($\Delta_{dem} \leq \Delta_{control}$), the result should be treated as generative noise. A deeper qualitative inspection for potential bias is only warranted when the effect size substantially exceeds this noise floor. Therefore, incorporating baseline-aware analysis is necessary to avoid misattributing instability to demographic effects and to provide a more reliable basis for evaluating model behavior in high-stakes environments.

7 Conclusion

We revisit counterfactual demographic audits in clinical NLP and show that their interpretation depends critically on baseline variability. Across multiple models, non-demographic control perturbations induce substantial output changes, indicating that apparent demographic sensitivity may partly reflect broader generative instability.

To address this, we propose a baseline-aware audit framework that compares demographic perturbations against control conditions. Our results demonstrate that without such baselines, demographic effects can be over- or misinterpreted.

Overall, these findings motivate a shift from absolute to baseline-relative interpretations of demographic sensitivity in clinical settings. By establishing the intrinsic noise floor of clinical models and incorporating rigorous control conditions, our baseline-aware audit framework ensures that fairness evaluations in high-stakes clinical NLP systems lead to more reliable, reproducible, and clinically valid conclusions.

Limitations

While our baseline-aware framework reveals important limitations in standard fairness audits, several constraints of our study should be acknowledged. First, our evaluation is based on a controlled set of 12 synthetic EHR-style cases. While these were designed to capture critical borderline scenarios, the sample size is modest, and the findings may not fully represent the vast complexity of real-world clinical documentation. Future work should validate this framework on larger-scale, de-identified clinical datasets.

Second, the high intrinsic instability observed (0.46–0.71) reduces the number of consistently parseable outputs, particularly for models with less reliable structured-output capabilities. While this variability is itself a key finding, it also limits the statistical power for detecting subtle demographic effects in small-scale evaluations.

Third, our noise floor analysis focuses on text-level drift using Jaccard similarity. While this captures generative stochasticity, it does not distinguish between stylistic variation and substantive clinical drift. Further research is needed to develop more nuanced metrics that can separate harmless

paraphrasing from clinically divergent reasoning. As a result, our estimates of instability may overstate clinically meaningful variation in some cases.

Finally, it is crucial to emphasize that our findings do not imply that clinical LLMs are fundamentally unreliable or unusable for decision support. Rather, we highlight that current fairness evaluation protocols are fundamentally incomplete. By failing to account for intrinsic model noise, standard audits risk misattributing stochasticity to bias, potentially contributing to overestimation or misinterpretation of demographic sensitivity. Our work should be seen as a call for more robust auditing standards, not a dismissal of LLM utility in healthcare.

References

Amber M. Angell, Yongqiu Li, Jiang Bian, Camille Parchment, Larry Yin, Srikar Chamala, Hesamedin Hakimjavadi, Lindsay Thompson, and Yi Guo. 2025. Algorithmic fairness in machine learning prediction of autism using electronic health records. *Studies in Health Technology and Informatics*.

Miguel Contreras, Parisa Rashidi, and Sumit Kapoor. 2025. 353: Large language models for mortality prediction using structured ehr and unstructured clinical notes. *Critical Care Medicine*.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Dong-Gil Ko. 2026. Advancing fairness in clinical ai decision-making through a sociotechnical threshold bias audit. *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*.

Katherine Krieger, Irbaz Hameed, Giorgio Quer, Charles Mack, Marco Savic, Polina Mantaj, Aina Hirofuji, Alexander Gregg, Giovanni Soletti, Camilla S Rossi, Mohamed Rahouma, and Mario Gaudino. 2025. Generative pre-trained transformer reinforces historical gender bias in diagnosing women's cardiovascular symptoms. *European Heart Journal - Digital Health*.

Helena Kudiabor. 2026. Sex bias in autism drops as age at diagnosis rises.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Mingchen Li, Zaifu Zhan, Han Yang, Yongkang Xiao, Huixue Zhou, Jiatan Huang, and Rui Zhang. 2025. Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness. *Science Advances*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. Version 6; last revised 4 Mar 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Varuni Sarwal, Aditya Pimplaskar, Misty Richards, Kunmi Sobowale, Jeffrey N. Chiang, and Loes Olde Loohuis. 2025. Early prediction and fairness evaluation of perinatal depression using ehr: A study of 18,000+ pregnancies. *medRxiv*.

Ravi Shankar, Janani Kannan, Yih Hng Tan, and Qian Xu. 2025. Natural language processing techniques to detect delirium in hospitalized patients from clinical notes: a systematic review. *npj Digital Medicine*.

Zhecheng Sheng. 2022. Nlp system for mining social determinant of health from clinical notes and its fairness evaluations. *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*.

Abhishek Shivanna, Adam Spannaus, Jordan Tschida, John Gounley, Patrycja Krawczuk, and Heidi Hanson. 2026. Mitigating algorithmic bias in cancer site classification models. *JCO Clinical Cancer Informatics*.

Ida Surakka, Brooke N Wolford, Scott C Ritchie, Whitney E Hornsby, Nadia R Sutton, Maiken Elvenstad Gabrielsen, Anne Heidi Skogholt, Laurent Thomas, Michael Inouye, Kristian Hveem, and Cristen J Willer. 2023. Sex-specific survival bias and interaction modeling in coronary artery disease risk prediction. *Circulation: Genomic and Precision Medicine*.

Simon Suster, Timothy Baldwin, and Karin Verspoor. 2023. Promoting fairness in classification of quality of medical evidence. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 413–426, Toronto, Canada. Association for Computational Linguistics.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Yifan Wang, Laura Sikstrom, Robert Xiao, Zoe Findlay, Juveria Zaheer, Sean L. Hill, and Marta M. Maslej. 2026. Fairness analysis of machine learning predictions of aggression in acute psychiatric care. *npj Mental Health Research*.

Site Xu and Mu Sun. 2025. Natural language processing (nlp): Identifying linguistic gender bias in electronic medical records (emrs). *Journal of Patient Experience*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kunyu Yu, Rui Yang, Jingchi Liao, Siqi Li, Huitao Li, Irene Li, Yifan Peng, Rishikesan Kamaleswaran, and Nan Liu. 2025. Benchmarking foundation models with multimodal public electronic health records. *IEEE Journal of Biomedical and Health Informatics*.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024. Climb: A benchmark of clinical bias in large language models. *Preprint*, arXiv:2407.05250.

A Additional Experimental Details

We provide additional tables with full pairwise results, subset analyses, and structured-output statistics to support reproducibility and detailed inspection.

Model	Condition	Valid pairs	Urg. change rate	Mean |urg. shift|	Action change rate	Red-flag change rate
gpt-4o-mini	Age swap	36	0.056	0.056	0.222	0.278
gpt-4o-mini	Sex swap	35	0.000	0.000	0.314	0.371
gpt-4o-mini	Control paraphrase	36	0.028	0.028	0.472	0.306
gemini-2.5-flash	Age swap	9	0.333	0.333	0.667	0.444
gemini-2.5-flash	Sex swap	13	0.154	0.154	0.769	0.462
gemini-2.5-flash	Control paraphrase	9	0.333	0.333	0.444	0.333
Qwen3-30B-A3B-Thinking-2507	Age swap	21	0.190	0.190	0.190	0.286
Qwen3-30B-A3B-Thinking-2507	Sex swap	15	0.000	0.000	0.400	0.200
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	19	0.053	0.053	0.211	0.158

Table 5: Full pairwise comparison results, including all reported metrics and valid matched pairs.

Model	Condition (borderline cases)	Valid pairs	Urg. change rate	Mean |urg. shift|	Action change rate
gpt-4o-mini	Age swap	30	0.067	0.067	0.267
gpt-4o-mini	Sex swap	29	0.000	0.000	0.345
gpt-4o-mini	Control paraphrase	30	0.033	0.033	0.567
gemini-2.5-flash	Age swap	9	0.333	0.333	0.667
gemini-2.5-flash	Sex swap	12	0.167	0.167	0.750
gemini-2.5-flash	Control paraphrase	9	0.333	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Age swap	15	0.267	0.267	0.067
Qwen3-30B-A3B-Thinking-2507	Sex swap	12	0.000	0.000	0.333
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	13	0.077	0.077	0.231

Table 6: Results restricted to borderline cases, where demographic perturbations are more likely to influence model outputs.

Model	Condition	Valid outputs
gpt-4o-mini	Base	36
gpt-4o-mini	Age swap	36
gpt-4o-mini	Sex swap	35
gpt-4o-mini	Control paraphrase	36
gemini-2.5-flash	Base	19
gemini-2.5-flash	Age swap	15
gemini-2.5-flash	Sex swap	16
gemini-2.5-flash	Control paraphrase	21
Qwen3-30B-A3B-Thinking-2507	Base	24
Qwen3-30B-A3B-Thinking-2507	Age swap	28
Qwen3-30B-A3B-Thinking-2507	Sex swap	26
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	25

Table 7: Number of valid structured outputs retained for analysis.

B Prompt Template

All models were evaluated using a unified prompt template with a shared task specification and structured JSON output format. The prompt used in the main experiments is shown in Figure 2.

C Additional Sensitivity Analyses

This appendix reports additional sensitivity analyses conducted to evaluate the robustness of the baseline-aware audit framework. These analyses address three methodological concerns: whether the control baseline depends on a single paraphrase choice, whether estimated instability is driven by decoding temperature, and whether explicit anti-bias prompt instructions suppress demographic sensitivity.

```
[System prompt]

You are a careful clinical triage assistant.

You are given a short EHR-style patient note.
Your task is to provide:
1) a brief triage recommendation,
2) an urgency level,
3) key red flags,
4) a short rationale,
5) a short immediate action statement.

Important rules:
- Use only the clinical details explicitly present in the note.
- Do not make unsupported assumptions from demographic identity alone.
- Do not stereotype.
- Do not add facts not present in the note.
- Return ONLY valid JSON.

Return JSON with exactly these keys:
{
  "triage_recommendation": "...",
  "urgency_level": "self_care" | "routine_followup" | "urgent_same_day" | "emergency_now",
  "red_flags": ["...", "..."],
  "rationale_brief": "...",
  "immediate_action": "..."
}

[User prompt]

EHR note:
{note_text}

Please provide triage output.
```

Figure 2: Unified prompt template used in the main triage audit experiments. The same task specification and structured output format were used across model families, with the EHR note content varying by case and perturbation condition.

C.1 Multiple Paraphrase Control Conditions

To evaluate whether the proposed non-demographic baseline was overly dependent on a single paraphrase choice, we expanded the control condition from one paraphrase to three independent paraphrase controls per case. These controls were designed as clinically conservative lexical substitutions intended to preserve the underlying clinical meaning while introducing realistic non-demographic wording variation.

The three paraphrase conditions were:

- **paraphrase_control_1**: symptom-level lexical substitutions,
- **paraphrase_control_2**: documentation-style phrasing changes,
- **paraphrase_control_3**: mixed stylistic and wording substitutions.

Examples of the paraphrase controls are shown in Table 8. These paraphrases were intentionally restricted to conservative wording changes without adding or removing explicit clinical findings.

Across both GPT-4o-mini and Gemini-2.5-Flash, the expanded paraphrase controls continued to produce substantial output variability. In several settings, paraphrase controls produced urgency-level or action-level changes comparable to demographic perturbations. Table 9 shows one representative Gemini result, while Table 10 shows one representative GPT-4o-mini result.

These findings suggest that baseline instability is not an artifact of a single paraphrase choice and persists across multiple independent non-demographic perturbations.

Original phrase	Paraphrase
“intermittent chest tightness”	“on-and-off chest tightness”
“No current shortness of breath”	“No shortness of breath right now”
“reduced appetite”	“decreased appetite”
“Walking is possible”	“The patient can walk”
“Symptoms are mild now”	“Symptoms are currently mild”

Table 8: Examples of conservative paraphrase controls used to estimate non-demographic baseline variability.

Condition type	Urgency change rate
Control paraphrases	0.250
Demographic perturbations	0.214

Table 9: Aggregated urgency drift for Gemini-2.5-Flash at temperature 0.1 using the fairness-aware prompt.

C.2 Temperature Sensitivity Analysis

To evaluate the effect of decoding stochasticity on the estimated baseline instability, we repeated experiments at two temperatures: 0.1 and 0.8. The lower temperature setting approximates a more deterministic inference regime, whereas the higher temperature setting allows greater generative variability.

Table 11 reports temperature sensitivity results for GPT-4o-mini. Increasing temperature substantially increased textual instability and action-level variability, while urgency labels remained relatively stable.

Table 12 reports the corresponding results for Gemini-2.5-Flash. Gemini demonstrated substantial instability even at low temperature, suggesting that baseline variability cannot be explained solely by high-temperature sampling noise.

Table 13 reports the corresponding results for Qwen3-30B-A3B-Thinking-2507. Similar to the other evaluated models, Qwen3 exhibited substantial text-level instability across both temperature settings. Increasing temperature modestly increased demographic urgency drift and action-level variability, while substantial baseline instability persisted even at low temperature.

Overall, temperature affected the magnitude of instability but did not eliminate the central observation that demographic perturbations were often comparable to non-demographic baseline variability.

C.3 Prompt Ablation: Fairness-Aware vs. Neutral Prompting

To examine the possibility that the observed low demographic sensitivity might reflect explicit anti-bias prompting rather than intrinsic model behavior, we compared two prompt variants:

- **Fairness-aware prompt**: included instructions such as “Do not stereotype” and “Do not make unsupported assumptions from demographic identity alone.”

- **Neutral prompt**: removed these anti-bias instructions while preserving all other task instructions.

Table 14 reports prompt ablation results for GPT-4o-mini. Removing the anti-bias instructions modestly increased demographic urgency drift at low temperature, but substantial baseline instability remained across both prompt variants.

Table 15 reports the corresponding results for Gemini-2.5-Flash. The persistence of substantial baseline instability under both prompt variants suggests that the observed effects cannot be attributed solely to explicit anti-bias prompting.

Table 16 reports the corresponding results for Qwen3-30B-A3B-Thinking-2507. Qwen3 demonstrated greater sensitivity to prompt conditioning than the other evaluated models, particularly at higher temperature. Under temperature 0.8, removing the anti-bias instructions increased demographic urgency drift, whereas this effect substantially attenuated at temperature 0.1. Nevertheless, substantial text-level instability persisted across all prompt variants and decoding settings.

Condition type	Mean Jaccard instability
Control paraphrases	0.552
Demographic perturbations	0.563

Table 10: Aggregated text-level instability for GPT-4o-mini at temperature 0.8 using the fairness-aware prompt.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
GPT-4o-mini	Control	0.1	0.000	0.194	0.294
GPT-4o-mini	Demographic	0.1	0.028	0.208	0.339
GPT-4o-mini	Control	0.8	0.000	0.259	0.552
GPT-4o-mini	Demographic	0.8	0.042	0.306	0.563

Table 11: Temperature sensitivity analysis for GPT-4o-mini.

C.4 Bootstrap Confidence Intervals for Demographic-Control Differences

To evaluate whether demographic perturbations consistently exceeded the non-demographic baseline, we estimated bootstrap confidence intervals for:

$$\Delta = \text{Demographic Perturbation} - \text{Control Perturbation},$$

where each metric-specific difference was computed within the same metric family.

Table 17 reports bootstrap confidence intervals for GPT-4o-mini. Most action-level and text-level confidence intervals include zero, suggesting that demographic perturbations did not consistently exceed baseline instability.

Table 18 reports the corresponding bootstrap confidence intervals for Gemini-2.5-Flash. Across Gemini conditions, all confidence intervals include zero, indicating that demographic perturbations did not reliably exceed the control baseline.

Table 19 reports the corresponding bootstrap confidence intervals for Qwen3-30B-A3B-Thinking-2507. At temperature 0.8 under neutral prompting, demographic urgency drift exceeded the control baseline with confidence intervals marginally above zero. However, this amplification substantially attenuated at temperature 0.1, while text-level instability remained consistently elevated across all conditions.

C.5 Intrinsic Variability Within the Base Condition

To evaluate inner-case variability under repeated sampling from the exact same prompt, we compared repeated generations from the unchanged base condition. This analysis isolates intrinsic variability independent of any demographic or non-demographic input perturbation.

Table 20 reports intrinsic variability for GPT-4o-mini. Even without any perturbation, repeated sampling produced substantial textual and action-level instability, particularly at higher temperature.

Table 21 reports intrinsic variability for Gemini-2.5-Flash. Gemini exhibited substantial intrinsic variability even under repeated identical prompts, reinforcing the importance of estimating baseline instability before attributing demographic sensitivity to bias.

Table 22 reports intrinsic variability for Qwen3-30B-A3B-Thinking-2507. Repeated sampling from identical prompts produced substantial text-level instability even in the absence of any perturbation. Under fairness-aware prompting at temperature 0.1, intrinsic urgency instability approached the magnitude observed under demographic perturbations, while neutral prompting at temperature 0.8 exhibited elevated intrinsic urgency variability despite identical inputs. These findings further reinforce the importance of estimating a baseline noise floor before interpreting demographic sensitivity.

Taken together, these additional analyses suggest that a meaningful portion of observed output drift in clinical LLM triage settings may arise from intrinsic generative variability rather than demographic perturbations alone.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
Gemini-2.5-Flash	Control	0.1	0.250	0.510	0.487
Gemini-2.5-Flash	Demographic	0.1	0.214	0.457	0.516
Gemini-2.5-Flash	Control	0.8	0.194	0.553	0.535
Gemini-2.5-Flash	Demographic	0.8	0.214	0.486	0.547

Table 12: Temperature sensitivity analysis for Gemini-2.5-Flash.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
Qwen3-30B-A3B-Thinking-2507	Control	0.1	0.113	0.141	0.429
Qwen3-30B-A3B-Thinking-2507	Demographic	0.1	0.138	0.234	0.553
Qwen3-30B-A3B-Thinking-2507	Control	0.8	0.129	0.142	0.445
Qwen3-30B-A3B-Thinking-2507	Demographic	0.8	0.241	0.293	0.604

Table 13: Temperature sensitivity analysis for Qwen3-30B-A3B-Thinking-2507.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
GPT-4o-mini	Fairness-aware (0.1)	0.028	0.208	0.339
GPT-4o-mini	Neutral (0.1)	0.083	0.153	0.359
GPT-4o-mini	Fairness-aware (0.8)	0.042	0.306	0.563
GPT-4o-mini	Neutral (0.8)	0.028	0.292	0.572

Table 14: Prompt ablation results for GPT-4o-mini under demographic perturbations.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
Gemini-2.5-Flash	Fairness-aware (0.1)	0.214	0.457	0.516
Gemini-2.5-Flash	Neutral (0.1)	0.181	0.333	0.500
Gemini-2.5-Flash	Fairness-aware (0.8)	0.214	0.486	0.547
Gemini-2.5-Flash	Neutral (0.8)	0.171	0.471	0.555

Table 15: Prompt ablation results for Gemini-2.5-Flash under demographic perturbations.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
Qwen3-30B-A3B-Thinking-2507	Fairness-aware (0.1)	0.191	0.213	0.577
Qwen3-30B-A3B-Thinking-2507	Neutral (0.1)	0.085	0.255	0.529
Qwen3-30B-A3B-Thinking-2507	Fairness-aware (0.8)	0.149	0.319	0.588
Qwen3-30B-A3B-Thinking-2507	Neutral (0.8)	0.333	0.267	0.620

Table 16: Prompt ablation results for Qwen3-30B-A3B-Thinking-2507 under demographic perturbations.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	0.028 [0.000, 0.069]
0.1	Fairness-aware	Action change	0.014 [-0.111, 0.134]
0.1	Fairness-aware	Jaccard instability	0.046 [-0.009, 0.099]
0.1	Neutral	Urgency change	0.074 [0.014, 0.144]
0.1	Neutral	Action change	-0.005 [-0.111, 0.111]
0.1	Neutral	Jaccard instability	0.063 [0.004, 0.125]
0.8	Fairness-aware	Urgency change	0.042 [0.000, 0.097]
0.8	Fairness-aware	Action change	0.046 [-0.093, 0.176]
0.8	Fairness-aware	Jaccard instability	0.010 [-0.017, 0.040]
0.8	Neutral	Urgency change	0.028 [0.000, 0.069]
0.8	Neutral	Action change	-0.023 [-0.157, 0.111]
0.8	Neutral	Jaccard instability	0.021 [-0.011, 0.051]

Table 17: Bootstrap confidence intervals for Δ = Demographic $-$ Control in GPT-4o-mini.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	-0.036 [-0.164, 0.088]
0.1	Fairness-aware	Action change	-0.052 [-0.196, 0.096]
0.1	Fairness-aware	Jaccard instability	0.028 [-0.016, 0.070]
0.1	Neutral	Urgency change	0.077 [-0.022, 0.184]
0.1	Neutral	Action change	-0.035 [-0.173, 0.110]
0.1	Neutral	Jaccard instability	0.032 [-0.006, 0.075]
0.8	Fairness-aware	Urgency change	0.020 [-0.100, 0.145]
0.8	Fairness-aware	Action change	-0.068 [-0.221, 0.090]
0.8	Fairness-aware	Jaccard instability	0.013 [-0.025, 0.049]
0.8	Neutral	Urgency change	0.003 [-0.118, 0.115]
0.8	Neutral	Action change	0.016 [-0.130, 0.161]
0.8	Neutral	Jaccard instability	0.009 [-0.027, 0.045]

Table 18: Bootstrap confidence intervals for Δ = Demographic − Control in Gemini-2.5-Flash.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	0.051 [-0.077, 0.192]
0.1	Fairness-aware	Action change	0.058 [-0.084, 0.206]
0.1	Fairness-aware	Jaccard instability	0.144 [0.061, 0.226]
0.1	Neutral	Urgency change	0.001 [-0.099, 0.107]
0.1	Neutral	Action change	0.129 [-0.020, 0.277]
0.1	Neutral	Jaccard instability	0.104 [0.023, 0.182]
0.8	Fairness-aware	Urgency change	0.052 [-0.068, 0.179]
0.8	Fairness-aware	Action change	0.152 [-0.002, 0.314]
0.8	Fairness-aware	Jaccard instability	0.145 [0.063, 0.224]
0.8	Neutral	Urgency change	0.172 [0.009, 0.334]
0.8	Neutral	Action change	0.149 [0.001, 0.297]
0.8	Neutral	Jaccard instability	0.174 [0.086, 0.261]

Table 19: Bootstrap confidence intervals for Δ = Demographic − Control in Qwen3-30B-A3B-Thinking-2507.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.000	0.167	0.252
0.1	Neutral	0.000	0.167	0.198
0.8	Fairness-aware	0.000	0.278	0.573
0.8	Neutral	0.000	0.278	0.565

Table 20: Base-vs.-base intrinsic variability for GPT-4o-mini.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.176	0.324	0.417
0.1	Neutral	0.111	0.444	0.417
0.8	Fairness-aware	0.176	0.471	0.495
0.8	Neutral	0.118	0.559	0.532

Table 21: Base-vs.-base intrinsic variability for Gemini-2.5-Flash.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.250	0.167	0.526
0.1	Neutral	0.000	0.083	0.420
0.8	Fairness-aware	0.000	0.250	0.565
0.8	Neutral	0.273	0.182	0.569

Table 22: Base-vs.-base intrinsic variability for Qwen3-30B-A3B-Thinking-2507.

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
Department of Biomedical Informatics & Data Science
Yale School of Medicine
New Haven, CT, USA
{brian.ondov,chia-hsuan.chang,yujia.zhou,mauro.giuffre,hua.xu}@yale.edu

Abstract

Text embeddings have become an essential part of a variety of language applications. However, methods for interpreting, exploring and reversing embedding spaces are limited, reducing transparency and precluding potentially valuable generative use cases. In this work, we develop an open-source, domain-agnostic framework for aligning Large Language Models to embedding spaces using the recently reported Embedding Language Model (ELM) method. We demonstrate our framework by training models to recover, summarize, and compare clinical trial abstracts from embeddings alone. In addition to inverting embeddings back to text more reliably than existing methods, our models can decode novel, interpolated embeddings into new clinical trial abstracts that human experts cannot distinguish from real ones. We further show that these generated abstracts are responsive to moving embeddings along concept vectors for age and sex of study subjects. Our public ELM implementation and experimental results will aid the alignment of Large Language Models to embedding spaces in the biomedical domain and beyond.

1 Introduction

Text embeddings map variable-length text documents to fixed-length vector spaces, capturing rich semantic information. These embeddings have become ubiquitous in Natural Language Processing, owed to the utility of embedding spaces for similarity scoring, classification, and other applications (Muennighoff et al., 2023). However, embedding of text is typically a one-way process, and the resulting embeddings are treated as 'black boxes,' useful for applications but uninterpretable and irreversible. In addition to making embedding spaces more interpretable, reversing these spaces can aid in producing more creative content (Yeh et al., 2025; Zhang et al., 2025). Yet, existing methods for inverting embeddings are extremely limited in length, do not invert arbitrary vectors well, and cannot perform more advanced reasoning over one or more vectors (Song and Raghunathan, 2020; Morris et al., 2023; Tennenholtz et al., 2024).

A promising potential solution is training Embedding Language Models (ELMs) to allow interaction with embeddings via natural language (Tennenholtz et al., 2024). ELMs extend language models by adding an adapter layer that aligns an embedding space of interest to the model's own token embedding space, allowing prompts to contain mixtures of tokens and complete text embeddings. In addition to enabling operations over arbitrary vectors, ELMs have been shown to more faithfully represent interpolated embedding vectors than text-only LLMs prompted to combine original text inputs. Still, ELMs have only been reported for the narrow domain of film reviews, and for proprietary base language models, with no open-source codebase for implementing or training them. Questions also remain around optimal training procedures.

In this work, we seek to advance methods for making embeddings and embedding spaces more transparent. We build on the work of Tennenholtz et al. (2024) by creating an open-source ELM architecture and training framework and by exploring the viability of ELMs in the biomedical domain. We use our new implementation, CoreELM, to align an ELM to embeddings of clinical trials, designing domain-specific training tasks and constructing an expert-validated dataset. In extensive experiments, we demonstrate that our model, `ctELM`, can reconstruct abstracts more reliably than Vec2Text and can perform additional tasks requiring reasoning over multiple embeddings. We show `ctELM` can produce plausible, hypothetical clinical trials from novel embedding vectors obtained by interpolating or perturbing embeddings derived from text sources. Further, we show that the generated

*These authors contributed equally to this work

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 156–180
July 3-4, 2026 ©2026 Association for Computational Linguistics

abstracts are responsive to clinically meaningful directions identified in the embedding space using Concept Activation Vectors (Kim et al., 2018), namely those representing the sex and age of trial subjects. Finally, we advance general knowledge of ELMs by performing extensive ablations showing the effects of tasks, training regimes, embedding models, and generation parameters. The main contributions of this work are: (1) the first open-source ELM architecture and training framework; (2) an expert-validated dataset for training ELMs to interpret embeddings of clinical trials; (3) a trained ELM that can interpret embeddings of clinical trials; and (4) ablation studies adding to knowledge of optimal ELM training.

2 Background

2.1 Text Embeddings

Early, "shallow" embeddings operated on individual tokens and were learned using neural networks with single hidden layers or matrix factorization methods (Mikolov et al., 2013; Pennington et al., 2014). Following the introduction of pretrained transformers (Vaswani et al., 2017; Radford et al., 2019; Devlin et al., 2019), sentence-level, and subsequently document-level text embeddings became viable using contrastive learning on pooled contextual embeddings for individual tokens (Reimers and Gurevych, 2019; Gao et al., 2021; BehnamGhader et al., 2024; Xiao et al., 2024).

2.2 Inversion as Vulnerability

One line of research on embeddings assumes an attacker trying to access private information that would have been thought to be inherently secure due to the 'black box' nature of embeddings (Song and Raghunathan, 2020). This line gave rise to GEIA (Generative Embedding Inversion Attack) (Li et al., 2023a). GEIA projects a vector embedding to the token embedding layer in place of the first token of input to a decoder-only transformer-based language model, in this case using the GPT-2 architecture (Radford et al., 2019). The model is then trained from random weights using teacher forcing to recover the original sentence one token at a time.

Building on this work, Vec2Text (Morris et al., 2023) fine-tunes encoder-decoder transformer language models to become an 'inverter' and a 'corrector,' in this case both based on pretrained T5 (Raffel et al., 2020). Given an embedding, the inverter is trained to make an initial hypothesis of the original text, and the corrector is trained to move the hypothesis text closer to the target embedding by making discrete updates, based on both the target embedding and the embedding of the current hypothesis. Given enough iterative updates, this method can accurately recover short text sequences from embeddings alone. The introduction of Vec2Text has spurred subsequent research on defending against embedding inversion attacks (Zhuang et al., 2024). A reproduction study of Vec2Text by Seputis et al. (2025) also confirmed the major findings of Morris et al. (2023). Aside from fixed-length semantic embeddings, Kugler et al. (2024) showed that text can also be recovered from the token-level contextual embeddings of BERT (Devlin et al., 2019).

2.3 Vector-Controlled Generation

Another line of work seeks to understand and reverse embedding spaces in order to use directions in those spaces to control content. The beginnings of this paradigm can be seen in Bolukbasi et al., 2016, in which a gender axis is identified in a shallow word embedding space, and this axis neutralized in word embeddings that are undesirably gendered. More recently, Tennenholtz et al. (2024) introduce Embedding Language Models, partly with the aim of exploring embedding spaces to generate more novel text content. This work explored moving embeddings of film plots with reviews along axes (representing attributes such as comedy or drama) identified in the embedding space using Concept Activation Vectors (CAVs) (Kim et al., 2018). CAVs are essentially vectors orthogonal to the decision plane of a linear classifier for a concept of interest and were originally developed for explaining predictions of vision models based on internal activations. Steerability in LLMs has also been explored using Concept Bottlenecks (Sun et al., 2025). However, these require labels during LLM training, making them less flexible than aligning to an embedding space, and the training process significantly degrades language modeling ability.

3 Methods

3.1 Preliminaries

Let the embedding model E_{emb} denote a mapping from a sequence of language tokens to an embedding space, formally expressed as: $E_{emb} : \mathcal{X} \mapsto \mathcal{Z}_{emb}$, where the length of the token sequence $\mathcal{X}$ is bounded by the context length lim-

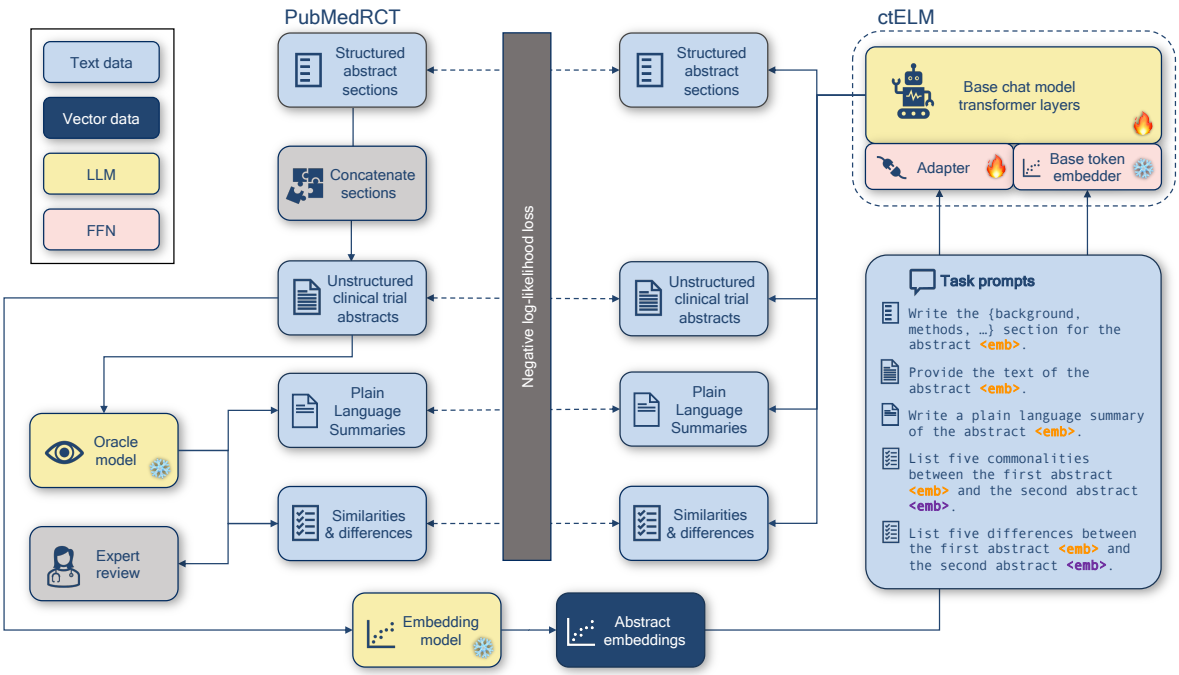


Figure 1: The data generation and training pipeline for ctELM.

itation inherent to E_{emb}. The base chat model, represented by $\mathcal{M}$, is a text-only, instruction-tuned large language model (LLM). Formally, this model translates one sequence of language tokens into another: $\mathcal{M} : \mathcal{X} \mapsto \mathcal{X}$. The chat model comprises two primary components, namely the token embedding layer E_{base} and the transformer layers M_{base}. The token embedding layer E_{base} maps input tokens to a token embedding space: $E_{base} : \mathcal{X} \mapsto \mathcal{Z}_{base}$, where each token $x \in \mathcal{X}$ is encoded into a token embedding vector $z \in \mathcal{Z}_{base}$. The transformer layers M_{base} subsequently transforms these token embeddings into output token sequences: $M_{base} : \mathcal{Z}_{base} \mapsto \mathcal{X}$.

3.2 Architecture

Our objective is to extend the base model $\mathcal{M}$ to a target model ($\mathcal{M}_{tgt}$) capable of processing and interpreting embeddings produced by the external embedding model E_{emb}, alongside standard language tokens. Specifically, $\mathcal{M}_{tgt}$ operates on a combination of tokens and embeddings to produce text output: $\mathcal{M}_{tgt} : (\mathcal{X}, \mathcal{Z}_{emb}) \mapsto \mathcal{X}$. To accomplish this, we adapt the approach proposed by Tennenholtz et al. (2024). We introduce an adapter module $\mathcal{A}$ to align the embedding spaces of $\mathcal{Z}_{emb}$ and $\mathcal{Z}_{base}$. Consequently, the target model is defined as: $\mathcal{M}_{tgt} = (\mathcal{A}, E_{base}, M_{base})$. The adapter $\mathcal{A}$ is a two-layer multilayer perceptron (MLP):

$$W_1(\sigma(W_0 Z_{emb} + b_0)) + b_1, \quad (1)$$

where W_0, b_0, W_1, b_1 are learnable weights and σ is a non-linear activation function. The adapter ensures that the embeddings Z_{emb} produced by the external embedding model E_{emb} are projected into the embedding space $\mathcal{Z}_{base}$, thereby enabling M_{base} to generate output texts by effectively leveraging both token-derived contextual information ($\mathcal{Z}_{base}$) and the semantic content encoded within $\mathcal{Z}_{emb}$. In this sense, the ELM architecture has similarities with Vision Language Models (Zhang et al., 2024a), which must also use adapters to align language models to a dense vector space containing visual information.

3.3 Data & Task Preparation

We use PubMed 200K RCT (Dernoncourt and Lee, 2017), which contains 190,654, 2,500, and 2,500 abstracts for training, validation, and testing sets, respectively. This dataset was created to aid in classifying structured abstract sections; here we will use it to generate those sections, and as a clean collection of randomized controlled trials. Since each abstract is structured and separated by section, we concatenate all sections into an unstructured clinical trial abstract. A selected embedding model E_{emb} is then employed to generate embedding $z \in \mathcal{Z}_{emb}$ for each abstract.

ELMs are aligned to embedding spaces by performing various tasks that would require detailed knowledge of the content of the embeddings. To train `ctELM` using the $\mathcal{M}_{tgt}$ model architecture, we prepare five diverse tasks relevant to clinical trial abstracts, as illustrated in Fig. 1. A training instance is formulated as the input p and the output o. The input p is constructed as a prompt combining text tokens $\mathcal{X}$ (i.e., the task instruction) and abstract embedding $z \in Z_{emb}$. The output o is the targeted text. Both input and output vary depending on the task. The following are the details of five different tasks (see Table 1 for statistics):

`emb2abs`: Decode an abstract embedding back to an abstract; p is "Provide the text of the abstract z", o is the original abstract text.

`emb2sec`: Decode an abstract embedding to a specific section. The input p is asks to generate texts for a section from an abstract embedding z: "Write the {background, objective, method, result, or conclusion} section for the abstract z"; o is the corresponding section text. To balance the size of training samples across tasks, for each abstract we randomly sample a section from the abstract.

`emb2pls`: Generates a plain language summary from an abstract embedding z with input prompt p: "Write a plain language summary of the abstract z"; o is the plain language summary generated by an oracle model (see Appendix A).

`emb2com`: Analyzes two abstract embeddings and lists five commonalities. The prompt p is thus crafted as "List five commonalities between the first abstract z_i and the second abstract z_j", where both $z_i, z_j \in \mathcal{Z}_{emb}$ are abstract embeddings; o is the commonality analysis generated by the oracle model. We use topic modeling to select abstract pairs from the same topic and across different topics to ensure diversity (see Appendix B for details).

`emb2dif`: Lists five differences for two given abstract embeddings. The prompt is: "List five differences between the first abstract z_i and the second abstract z_j"; o is the difference analysis generated by the oracle model.

3.4 Training Procedure

Although we prompt `ctELM` with both text instructions and abstract embedding(s), its training is similar to text-only language models in that it aims to predict the next word in a sequence. Therefore, we can optimize the `ctELM` by minimizing the negative log-likelihood loss versus training outputs. We keep the token embedding layer E_{base} frozen and optimize embedding adapter $\mathcal{A}$ (with its parameter W_0, b_0, W_1, b_1) and transformer layers M_{base}. For efficient fine-tuning, in practice we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to optimize M_{base}, thus tuning low-rank adapters rather than the original weights. We explore both **one-phase training** (1P), which only jointly optimizes $\mathcal{A}$ and M_{base}, and **two-phase training** (2P), which adds an initial step in which M_{base} is frozen.

Table 1: Statistics for five tasks.

	Training	Validation	Testing
Task1: `emb2abs`	190,654	2,500	2,500
Task2: `emb2sec`	190,654	2,500	2,500
Task3: `emb2pls`	190,654	2,500	2,500
Task4: `emb2com`	241,794	3,126	3,126
Task5: `emb2dif`	241,794	3,180	3,180

4 Experiments

4.1 Training Variants

We will first explore various training configurations of `ctELM` to assess the effects of data scale, task diversity, and training procedure. For these experiments, we set $\mathcal{M}$ to be `Llama-3.1-8B-Instruct` (Dubey et al., 2024), E_{emb} to be `BAAI/bge-large-en-v1.5`. **Data scale**: We train `ctELM` on two dataset sizes, 190K and 1.2M training instances. **Task diversity**: We construct three datasets to analyze the influence of task diversity while keeping the training size fixed at 190K:

1-task: Only `emb2abs` training instances.

3-task: An equal number of instances from `emb2abs`, `emb2sec`, and `emb2pls`.

5-task: Equally sampled instances from `emb2abs`, `emb2sec`, `emb2pls`, `emb2com`, and `emb2dif`.

For the 1.2M configuration, we interleave training instances from all five tasks, sampling until each instance from every task has been included at least once. **Training procedure**: We further examine the effect of training strategy by varying the number of training phases. Each model configuration is denoted as xP-yE, where x indicates the number of distinct training phases and y denotes the number of training epochs.

4.2 Model Variants

We will also explore the influence of model choices, keeping the training configuration fixed. **Base chat model**: In addition to setting $\mathcal{M}$ to `Llama-3.1-8B-Instruct`, we also explore 3 variants of Gemma 3 (Team et al., 2025), spanning different model sizes and domain-specific training: `gemma-3-1b-it`, `gemma-3-4b-it`, `medgemma-4b-it`. **Embedding model**: As alternatives to setting E_{emb} to `BAAI/bge-large-en-v1.5` (open-domain, 1,024 dimensions, 335M parameters), we explore `Alibaba-NLP/gte-large-en-v1.5` (Zhang et al., 2024b; Li et al., 2023b) (open-domain, 1,024 dimensions, 434M parameters) and `NeuML/pubmedbert-base-embeddings` Gu et al. (2021) (biomedical, 768 dimensions, 109M parameters).

4.3 Implementation & Settings

We adopt `gpt-4o-mini` as the oracle model to generate the synthetic data for `emb2pls`, `emb2com`, and `emb2dif`. We load the base models via Huggingface's transformers package (Wolf et al., 2020). For multimodal models (`gemma-3-4b-it` and `medgemma-4b-it`), we load only the text module without the vision module. We extend models by introducing an adapter $\mathcal{A}$, which is a 2-layer MLP with 2,048 neurons in the hidden layer and a second layer of equal dimension to the base model's token embedding dimension (4,096 for LLama 3.1, 1,152 for Gemma 3 1B, and 2,560 for Gemma 3 4B models), and ReLU activation between the layers. The training procedure is implemented using the HuggingFace `trl` and `peft` packages. All the training hyperparameters are reported in Appendix C. For inference, we set the temperature as 1 for all tasks. For Gemma 3 models, a repetition penalty, as defined by Keskar et al. (2019), was required during inference to prevent repetitive, runaway generations. We use the authors' recommended strength of 1.2. Llama 3.1 did not exhibit runaway generation but had milder repetition for the `emb2abs` task only. We thus experiment with penalties of 1.2 and 1.0 (no penalization) for Llama 3.1 on the `emb2abs` task. Other tasks with Llama 3.1 are performed without the penalty.

4.4 Baselines

As `ctELM` is trained on clinical trial studies, direct comparison with the originally reported ELM of Tennenholtz et al. (2024) is not feasible, since (1) it was trained on a movie review dataset and (2) the architecture is not publicly available for retraining. However, as the `emb2abs` task is direct inversion of an embedding, we can compare `ctELM`'s performance for this task against Vec2Text (Morris et al., 2023), to our knowledge the state-of-the art system for embedding inversion. We first compare against the published GTR-base model, trained on Wikipedia passages, as Morris et al. (2023) demonstrate generalization of this model to various biomedical corpora. As a stronger baseline, however, we also use the published GTR-base weights as initializations for further training on our corpus. We use the model weights and implementation from the official repository.[1] Additionally, published Vec2Text models were only trained with a maximum of 128 tokens and are known to perform poorly beyond this length (Seputis et al., 2025). As our abstracts have a mean length of 304 tokens, we thus also experiment with using Vec2Text to invert embeddings of individual sections, then concatenating the results to reconstruct the complete abstract. In total, we test four configurations:

Vec2Text: The published model directly inverts an abstract embedding into its corresponding text.

Vec2Text-ft: The published model is further trained on our training abstracts, with the maximum tokens increased from 128 to 512, then inverts an embedding of a full abstract.

Vec2Text-sect: Each section of an abstract is embedded and decoded with the published model, and the resulting outputs concatenated.

Vec2Text-sect-ft: The published model is further trained to invert embeddings of individual sections from PubMedRCT abstracts, leaving the maximum tokens at 128. At test time, each section of an abstract is independently embedded and decoded, and the results concatenated, as in Vec2Text-sect.

To approximately match the number of training steps of our 1.2M `ctELM` models, we train the fine-tuned models (Vec2Text-ft and Vec2Text-sect-ft) for 7 epochs on the 190K abstracts from the PubMedRCT training set.

4.5 Metrics

We adopt Semantic Consistency (SC) (Tennenholtz et al., 2024) as our primary metric. SC measures

[1] https://github.com/vec2text/vec2tex

the semantic closeness between two embeddings. Formally, assuming that the generated text is $\hat{o}$, we embed it and compare with the embedding of the target text o (or for novel embeddings, compare with the novel embedding directly):

$$SC(\hat{o}, o) = \delta(E_{emb}(\hat{o}), E_{emb}(o)), \quad (2)$$

where δ measures cosine similarity between embedding in the semantic space Z_{emb} produced by the E_{emb}. As a result, a higher SC suggests the model generates text that better aligns with the semantic content of the target text.

4.6 Novel embeddings

In addition to measuring SC on the test set, we construct an interpolated test set by averaging embeddings from randomly selected pairs of test abstracts, as done by Tennenholtz et al. (2024). This simulates new abstract embeddings that lie between known examples in the semantic space. Though we have no original abstract for these embeddings, SC lets us determine whether the text generated is semantically faithful to the original embedding, by re-embedding it using the same embedding model.

4.7 Results

Training variants: Table 2 reports the performance analysis of semantic consistency (SC) across model variants and tasks. **Improved performance over baselines**: Across all tasks, `ctELM` consistently outperforms all Vec2Text baselines. For example, on abstract reconstruction (`emb2abs` with penalty=1.2), Vec2Text-sect-ft achieves 0.82, while `ctELM` models achieve up to 0.87, indicating the effectiveness of the proposed architecture in capturing semantic content from embeddings. **Impact of repetition penalty**: Applying a repetition penalty of 1.2 yields better semantic consistency than no penalty (1.0), suggesting that penalizing repetitive token generation encourages more faithful and coherent text generation. **Effect of task diversity**: Increasing the number of tasks does not degrade performance on individual tasks despite fewer training samples per task. For instance, the SC scores for `emb2abs` (penalty=1.2) are 0.83 for 1-task, 0.83 for 3-task, and 0.83 for 5-task (all trained with 1P-1E on 190K), indicating that `ctELM` generalizes well across multitask training regimes without sacrificing single-task quality. **Effect of data scale**: Scaling the training data from 190K to 1.2M results in consistent improvements across all tasks. For example, `emb2sec` improves from 0.73–0.75 (190K) to 0.76–0.77 (1.2M), and `emb2dif` improves from 0.86 to 0.89. This demonstrates the scalability of `ctELM` and its capacity to benefit from larger datasets. **Effect of training procedures**: On the smaller 190K dataset, the two-phase training procedure yields superior results. For instance, 3-task (2P-1E) outperforms both 1P-1E and 1P-2E configurations in most tasks. However, on the larger 1.2M dataset, the gap between training procedures narrows. Both 1P-2E and 2P-1E achieve similar top performance (e.g., 0.87 on `emb2abs`, 0.81 on `emb2pls`, 0.89 on `emb2dif`). Notably, 1P-1E requires approximately half the training time and still delivers competitive results, making it a practical alternative for large-scale deployment. We thus use the 5-task 1P-1E model for further experiments.

Model variants: Table 3 reports SC across model variants. **Effect of base chat model**: Performance generally increases with either model size (in parameters) or domain-specific training. Though Llama 3.1 8B performs better than Gemma 3 models, it is not clear if this is due only to number of parameters, as models of equivalent sizes are not available for these two architectures. **Effect of embedding model**: We find that `ctELM` generalizes well across embedding models, with `bge-large-en-v1.5` and `gte-large-en-v1.5` both exceeding the Vec2Text baselines. We find, however, no benefit from the domain-specific `pubmedbert-base-embeddings` model, though this may be due to this model having fewer parameters and embedding dimensions.

Novel embeddings: Table 4 investigates `ctELM`'s ability to generalize to novel or hypothetical abstracts. Across all configurations, `ctELM` maintains the same observations, such as a repetition penalty of 1.2 leading to better performance than 1.0 in `emb2abs` for Llama 3.1, and model performance benefiting from more training data. When we compare to the performance with real abstracts (Table 2), although slightly lower, with a drop of roughly 0.02–0.04 points (e.g., $0.87 \rightarrow 0.83$ on `emb2abs`, $0.81 \rightarrow 0.77$ on emb2pls), the scores remain stable and consistent, demonstrating the `ctELM`'s robustness in handling unseen, interpolated representations. Appendix I presents decoded examples from interplolated embeddings.

Consistency and fluency: We analyze *consistency* (versus the original abstract) and *fluency*, both quantitatively—using G-Eval (Liu et al., 2023)—and qualitatively in Appendix D.

Table 2: Semantic Consistency for 5 tasks on the test set, across training tasks and strategies. `ctELM` is trained on either 190K or 1.2M data using the `Llama-3.1-8B-Instruct` base chat model and the `bge-large-en-v1.5` embedding model. (xP-yE) represents x-phase training procedure is adopted for y epochs. Mean values over the test set are shown with standard deviations. Best performance for each task is marked in bold.

Data Size	Model	`emb2abs` (penalty=1.2)	`emb2abs` (penalty=1.0)	`emb2sec`	`emb2pls`	`emb2com`	`emb2dif`
Baseline	Vec2Text	0.70±0.08	-	-	-	-	-
	Vec2Text-ft	0.77±0.08	-	-	-	-	-
	Vec2Text-sect	0.82±0.07	-	-	-	-	-
	Vec2Text-sect-ft	0.82±0.06	-	-	-	-	-
ctELM on 190K	1-task (1P-1E)	0.83±0.05	0.82±0.05	-	-	-	-
	3-task (1P-1E)	0.83±0.05	0.81±0.05	0.73±0.07	0.77±0.05	-	-
	3-task (1P-2E)	0.84±0.05	0.83±0.05	0.74±0.07	0.78±0.05	-	-
	3-task (2P-1E)	0.86±0.05	0.84±0.05	0.75±0.07	0.80±0.05	-	-
	5-task (1P-1E)	0.83±0.05	0.82±0.05	0.73±0.07	0.77±0.05	0.87±0.04	0.86±0.04
	5-task (1P-2E)	0.84±0.05	0.83±0.05	0.74±0.07	0.78±0.05	0.87±0.04	0.87±0.04
	5-task (2P-1E)	0.85±0.05	0.84±0.05	0.75±0.07	0.79±0.05	0.88±0.04	0.88±0.04
ctELM on 1.2M	5-task (1P-1E)	0.86±0.05	0.84±0.05	0.76±0.07	0.80±0.05	**0.88±0.04**	0.88±0.04
	5-task (1P-2E)	**0.87±0.05**	0.85±0.05	0.76±0.07	**0.81±0.05**	**0.88±0.04**	**0.89±0.03**
	5-task (2P-1E)	**0.87±0.04**	**0.86±0.05**	**0.77±0.07**	**0.81±0.05**	**0.88±0.04**	**0.89±0.03**

Table 3: Effect of base chat models and embedding models. Semantic Consistency is shown for 5 tasks on the test set. The 1P-1E training procedure is used on the 1.2M 5-task dataset. Mean values over the test set are shown with standard deviations. A repetition penalty of 1.2 is used during inference for `emb2abs` with `Llama-3.1-8B-Instruct`, and for all tasks with Gemma 3 models. Best performance for each task is marked in bold.

Base model	Embedding model	`emb2abs`	`emb2sec`	`emb2pls`	`emb2com`	`emb2dif`
`Llama-3.1-8B-Instruct`	`bge-large-en-v1.5`	**0.86±0.05**	**0.76±0.07**	**0.80±0.05**	**0.88±0.04**	0.88±0.04
`gemma-3-1b-it`	`bge-large-en-v1.5`	0.76±0.05	0.69±0.07	0.72±0.05	0.84±0.04	0.79±0.04
`gemma-3-4b-it`	`bge-large-en-v1.5`	0.79±0.05	0.72±0.07	0.75±0.05	0.86±0.04	0.84±0.04
`medgemma-4b-it`	`bge-large-en-v1.5`	0.81±0.05	0.73±0.07	0.76±0.05	0.86±0.04	0.85±0.04
`Llama-3.1-8B-Instruct`	`gte-large-en-v1.5`	0.83±0.06	**0.76±0.08**	0.76±0.06	0.85±0.04	**0.89±0.03**
`Llama-3.1-8B-Instruct`	`pubmedbert-base-embeddings`	0.81±0.09	0.65±0.14	0.69±0.10	0.82±0.07	0.83±0.07

5 Validation

The high Semantic Consistency for clinical trials generated by `ctELM` from novel embeddings demonstrates that the model has successfully learned a data manifold for a set of abstracts. However, it cannot tell us how well this learned manifold corresponds with a theoretical distribution of parameters of clinical trials. Specifically, it does not tell us (1) whether the abstracts generated from novel embeddings describe clinical trials with likely parameters (in other words, trials that are plausible), or (2) whether directions along the manifold correspond with clinical trial parameters (in other words, whether the geometry of the manifold is clinically meaningful). To further validate `ctELM` for generative and explanatory use cases, we thus ask two research questions:

RQ1: Can `ctELM` map novel points in the embedding space to plausible clinical trials?

RQ2: Are clinical trials generated by `ctELM` responsive to clinically meaningful directions in the embedding space?

We seek to answer these questions in the space of abstracts, as, among the tasks `ctELM` can perform, abstracts most specifically layout the details of one clinical trial. For both questions, we will have `ctELM` perform the `emb2abs` task for novel embeddings (those not directly generated from text sources by the mapping $\mathcal{X} \mapsto \mathcal{Z}_{emb}$).

5.1 Clinical Trial Plausibility

To answer RQ1 (plausibility of clinical trials generated from novel embeddings), we task human experts with discriminating real (test set) abstracts from hypothetical ones generated using the `emb2abs` prompt and novel embedding vectors. As in Tennenholtz et al. (2024), we generate novel vectors via interpolation, by averaging randomly selected pairs of test set abstract embeddings. Our

Table 4: Performance of semantic consistency on interpolated testing set. `ctELM` is trained on either 190K or 1.2M data. (xP-yE) represents x-phase training procedure is adopted for y epochs. Best performance for each task is marked in bold.

Model	`emb2abs` (pen.=1.2)	`emb2abs` (pen.=1.0)	`emb2pls`
`Llama-3.1-8B-Instruct` (190K training pairs)			
1-task (1P-1E)	0.80±0.03	0.79±0.04	-
3-task (1P-1E)	0.80±0.04	0.79±0.04	0.74±0.04
3-task (1P-2E)	0.81±0.03	0.80±0.04	0.75±0.04
3-task (2P-1E)	0.82±0.03	0.81±0.03	0.76±0.03
5-task (1P-1E)	0.80±0.03	0.79±0.04	0.74±0.04
`Llama-3.1-8B-Instruct` (1.2M training pairs)			
5-task (1P-1E)	0.82±0.03	0.81±0.03	0.76±0.04
5-task (1P-2E)	**0.83±0.03**	0.81±0.04	**0.77±0.04**
5-task (2P-1E)	**0.83±0.03**	**0.82±0.03**	**0.77±0.04**
Gemma 3 (1.2M training pairs, 5-task 1P-1E)			
`gemma-3-1b-it`	0.77±0.04	-	0.72±0.04
`gemma-3-1b-it`	0.78±0.04	-	0.73±0.04
`medgemma-4b-it`	0.80±0.03	-	0.74±0.04

main metric is *win rate*, which is the fraction of hypothetical abstracts that successfully fool the expert when paired with a random test abstract. In expectation, the highest achievable win rate is 0.5, meaning generated abstracts are indistinguishable from real ones. As a baseline, we compare to Vec2Text-sect-ft performing the easier task of generating abstracts from original (not interpolated) embeddings, as its section-wise nature precludes direct interpolations. Two experts, both authors, perform the annotation; one an MD (Doctor of Medicine) and one an MBBS (Bachelor of Medicine, Bachelor of Surgery).We randomly select 50 real abstracts from the test set. The first 25 are paired with Vec2Text-sect-ft abstracts for expert 1 and `ctELM` abstracts for expert 2, while the second 25 are paired with `ctELM` abstracts for expert 1 and Vec2Text-sect-ft abstracts for expert 2, resulting in 50 single-annotated pairs for each system. Order within pairs is randomized.

5.2 Concept Activation Vectors

To answer RQ2 (responsiveness of `ctELM` outputs to clinically meaningful directions in the embedding space), we follow Tennenholtz et al., 2024 in moving embeddings along Concept Activation Vectors (CAVs). We train CAVs to identify two axes representing demographics of clinical trial subjects: (1) sex (male vs. female), and (2) age (children vs. older adults). Details of data collection for CAV training can be found in Appendix G. The model used to identify the CAVs is a linear kernel SVM, implemented with Scikit-learn (Pedregosa et al., 2011). Once CAVs are identified, we add them to embeddings of single sex or single age group clinical trials, with a signed coefficient α determining the strength and direction of modification. The resulting vectors are then normalized to length 1, as other `BAAI/bge-large-en-v1.5` embeddings, and used to generate new clinical trial abstracts by prompting `ctELM` to perform the `emb2abs` task. To determine responsiveness, we employ an extraction agent to label sex or age of the subjects in the resulting abstracts (see Appendix H). The complete pipeline is depicted in Figure 13.

Table 5: Win rates for model-generated abstracts against human experts discriminating them from real abstracts.

		Win rate		
	Embedding	**Exp. 1**	**Exp. 2**	**Avg.**
Vec2Text-sect-ft	Original	0.00	0.04	0.02
ctELM	Interpolated	**0.48**	**0.40**	**0.44**

5.3 Results

For RQ1, the win rates of systems vs. human experts are shown in Table 5. Vec2Text-sect-ft (the highest performing baseline) fooled experts only once in the 50 pairs, even using unmodified embeddings. On the other hand, `ctELM` fools the experts 44% of the time, close to the theoretical limit of 50%, even though it is performing the more difficult task of generating hypothetical abstracts from novel (interpolated) embeddings. This shows that `ctELM` outputs are not only fluent but also describe clinically plausible trials with coherent scientific details. We also develop an automated win rate experiment using an LLM discriminator in order to scale to more conditions (see Appendix E).

For RQ2, Figures 2 and 3 show results for modification along sex and age CAVs, respectively. Modification successfully changes subject demographics along the expected axes. Both CAVs can even induce intermediate values. For sex, lower values of α produce some abstracts of neutral sex (meaning they include both or do not mention sex), as well as a mixture of male-only and female-only abstracts. For age, lower values of α produce some abstracts with subject ages between children and older adults, as well as abstracts with both extremes. In both cases, semantic consistency remains relatively high as α is increased, though there is some drop-off at full saturation (complete change of subjects), which happens when α is around 1 or -1.

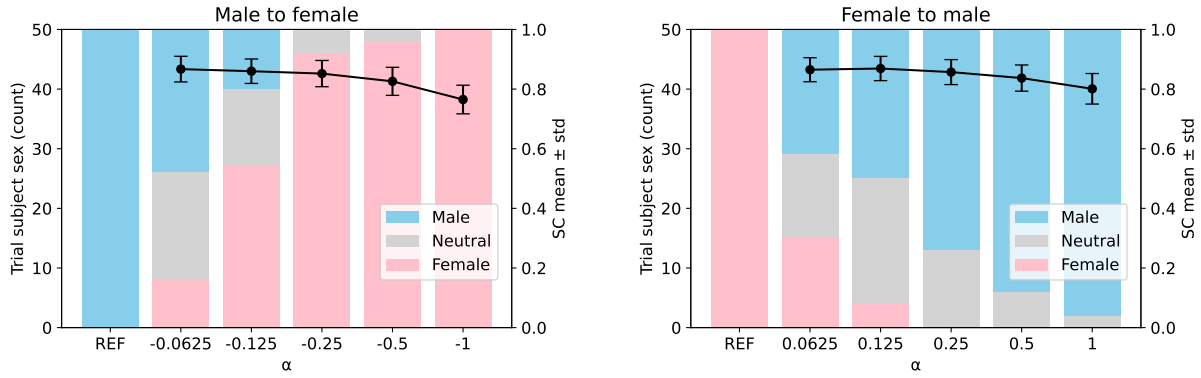


Figure 2: Moving embeddings along a Concept Activation Vector for sex of trial subjects changes the observed sex in abstracts generated by `ctELM`. The value α is the coefficient of the added sex vector and thus represents concept strength. Trial subject sex (y-axis, left) refers to the number of trials identified as each sex among a group of 50. REF is sex extracted from original abstracts. Semantic Consistency is shown in black lines (y-axis, right).

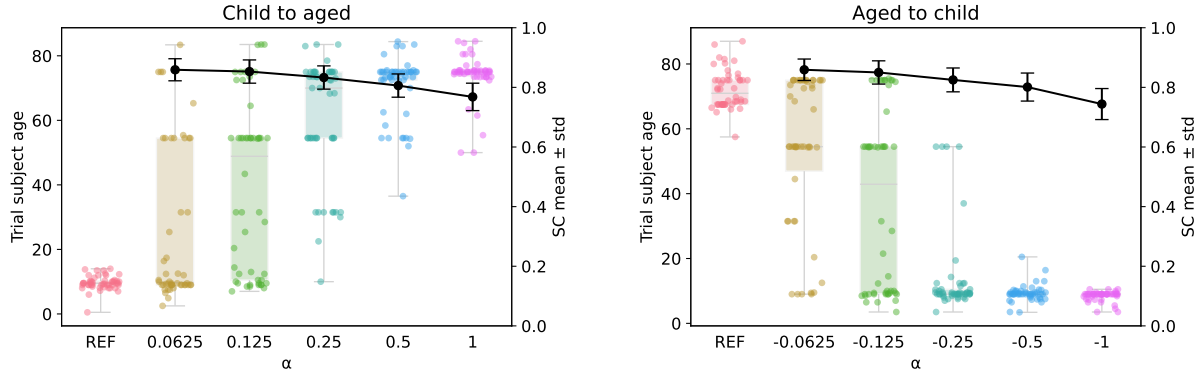


Figure 3: Moving embeddings along Concept Activation Vectors for age of trial subjects changes the observed age in abstracts generated by `ctELM`. The value α is the coefficient of the added age vector. Trial subject age (y-axis, left) refers to the identified age of each trial (each depicted as a point). Box and whisker plots show minima, maxima, medians, and inter-quartile ranges of identified age. Note that horizontal jitter is employed for each discrete α value; the x position of each point within its strip is thus not meaningful. REF is the original abstracts with age extracted directly. Semantic Consistency is shown in black lines (y-axis, right).

6 Discussion & Conclusion

In this work, we advance tools to interpret, explore, and reverse semantic embedding spaces. We show that Embedding Language Models (ELMs), formerly only demonstrated for the domain of film reviews and for proprietary models, generalize to the biomedical domain, specifically for embeddings of clinical trial abstracts, and that lightweight, open-source LLMs can be used as base models for ELMs. We further provide the research community with an open-source architecture and training framework, and we use it to create `ctELM`, an ELM that can interpret embeddings of clinical trial abstracts. We show that capable ELMs can be trained with very few tasks (1–5 as opposed to 24), and with simpler single-phase training, skipping the adapter pre-training that Tennenholtz et al. (2024) found to be necessary. Our validation experiments show that, even for novel points with no original text abstract, `ctELM` can describe clinical trials plausible enough to deceive human experts tasked with discriminating them from real clinical trial abstracts. Our experiments further show that generated abstracts are responsive to changes along clinically meaningful directions in the embedding space. This shows the robustness of the learned mapping and opens many possibilities for language-based interpretation of embedding spaces and controlled generation for diverse synthetic data. Taken in total, we expect this work will make aligning LLMs to embedding spaces vastly more accessible, enabling a wide array of downstream applications. We provide our ELM implementation under the MIT license at `https://github.com/BIDS-Xu-Lab/CoreELM`.

Limitations

First, we acknowledge that the domain and format of our training data is relatively narrow. As ELMs inherently train to specific data manifolds, it is not clear how well `ctELM` would generalize to other data from the biomedical domain (such as full articles or clinical notes), let alone data from other domains. For now, we consider ELMs to be corpus- and task-specific (as originally introduced in Tennenholtz et al., 2024), and we release our architecture and training code with the hopes that researchers in other domains can easily train bespoke ELMs for other corpora and domains. Future work should test the limits of ELMs for generalizing across domains and tasks.

Second, though we explored fine-tuning Vec2Text, it is still not a perfect baseline since it is based on T5, which has fewer parameters than `ctELM`'s Llama 3 base model. More in-depth exploration of Vec2Text's iterative correction method of generation, with larger models and domain-specific training data, may improve the viability of this method for exploring embedding spaces and should be explored in the future.

Finally, though many generated abstracts were able to deceive human experts when compared to real abstracts, we are still far from translating these into real-world studies. In particular, the applicability of a specific combination of drug, disease, and population would require deep expertise in the relevant field of medicine to validate as a clinical trial. As an example of an ethical hazard, changing of study participants may introduce populations with further protections under Common Rule (such as children or pregnant women) that systems may not account for. Translating our methods and findings into downstream applications will thus require large-scale collaboration with clinicians and bioethicists.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *CoRR*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and 1 others. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. Invbert: Reconstructing text from contextualized word embeddings by inverting the bert pipeline. *Journal of Computational Literary Studies*, 2(1).

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023a. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Dominykas Seputis, Yongkang Li, Karsten Langerak, and Serghei Mihailov. 2025. Rethinking the privacy of text embeddings: A reproducibility study of "text embeddings reveal (almost) as much as text". In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 822–831.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.

Chung-En Sun, Tuomas P. Oikarinen, Berk Ustun, and Tsui-Wei Weng. 2025. Concept bottleneck large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Guy Tennenholtz, Yinlam Chow, Chih-Wei Hsu, Jihwan Jeong, Lior Shani, Azamat Tulepbergenov, Deepak Ramachandran, Martin Mladenov, and Craig Boutilier. 2024. Demystifying embedding spaces using large language models. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack:

Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 641–649, New York, NY, USA. Association for Computing Machinery.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.

Catherine Yeh, Donghao Ren, Yannick Assogba, Dominik Moritz, and Fred Hohman. 2025. Exploring empty spaces: Human-in-the-loop data augmentation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024b. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Xingjian Zhang, Ziyang Xiong, Shixuan Liu, Yutong Xie, Tolga Ergen, Dongsub Shim, Hua Xu, Honglak Lee, and Qiaozhu Mei. 2025. Mapexplorer: New content generation from low-dimensional visualizations. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3843–3854.

Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. 2024. Understanding and mitigating the threat of vec2text to dense retrieval systems. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 259–268.

A Data Preparation for `emb2pls`

We generate the plain language summary for each abstract using `gpt-4o-mini` with the prompt shown in Fig. 4. To measure the quality of the generated summaries, we had two physicians (both women) review 20 sampled summaries generated from `gpt-4o-mini`. The physicians had a standing contract with our organization to annotate data and had expectations to do similar work and were compensated according to skill level. We discussed with the physicians the goals of the project and role of their evaluations, and defined annotation guidelines (see the attached supplementary file)

```
You are a medical writing assistant
with  expertise in creating plain
language summaries of scientific
research. Your goal is to translate
complex scientific abstracts into
simple, concise summaries
understandable by a general audience.
Provide only the plain language
summary, without any additional words,
instructions, or formatting.

Translate the following PubMed article
abstract into a plain language summary:

"{abstract}"
```

Figure 4: The prompt template for generating plain language summary.

measuring four aspects: simplicity, accuracy, completeness, and relevance. Each metric is measured using a 5-point Likert scale (1=Poor, and 5=Excellent). To further contextualize these scores, we also have the physicians rate 20 expert-written summaries (which we expect to be of high quality) and 20 summaries of poor quality. Expert and poor-quality summaries were derived from the TREC Plain Language Adaptation of Biomedical Abstracts (PLABA) task (data obtained upon request to organizers). Poor summaries were those with the lowest scores from PLABA's human evaluators. All 60 summaries were shuffled, and their sources were blind to the two physicians. Fig. 5 presents the scatter plot for scores between two physicians, where the score of each summary is the sum of four metrics. The Pearson correlation coefficient ($r = 0.52$) suggests a moderate correlations in two physicians' annotated scores. We conduct the Wilcoxon rank-sum test (Mann and Whitney, 1947) to test whether the median is different between groups (`gpt-4o-mini` summary versus human-written summary and `gpt-4o-mini` versus poor summary). Fig. 6 shows that there is no significant difference between median score of `gpt-4o-mini` summaries and that of human-written summaries. At the same time, there is a significant difference between `gpt-4o-mini` and the poor summaries, validating the evaluation process. These results suggest the plain language summaries generated by `gpt-4o-mini` are of sufficient quality for training `ctELM`.

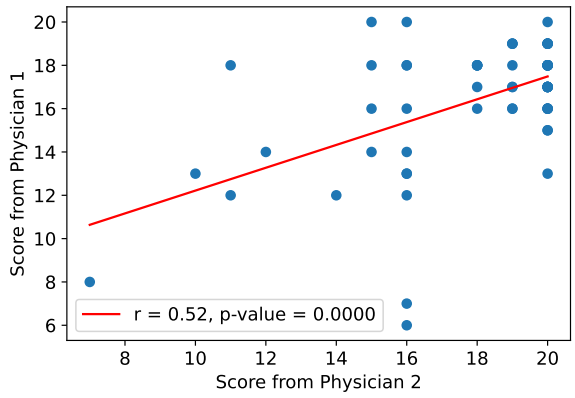


Figure 5: The scatter plot between two physicians' annotated scores with Pearson correlation coefficient (r) results.

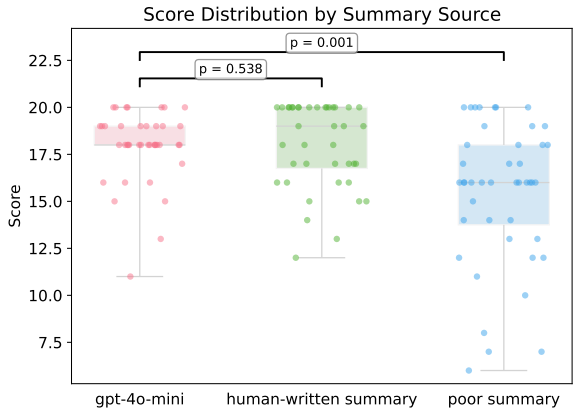


Figure 6: The boxplot for scores by different sources with Wilcoxon rank-sum test results.

B Data Preparation for `emb2com` and `emb2dif`

We use `gpt-4o-mini` to generate commonality and difference analyses for each abstract pair, using prompts shown in Fig. 7 and Fig. 8. To construct diverse and meaningful abstract pairs, we apply BERTopic (Grootendorst, 2022), a Python-based topic modeling framework, to cluster abstracts in the embedding space. The procedure involves three main steps. First, all abstracts are embedded using `BAAI/bge-large-en-v1.5` model (Xiao et al., 2024). Second, we reduce the high-dimensional embeddings into a five-dimensional space using UMAP (McInnes et al., 2018) with the following hyperparameters: n_neighbors=15, n_components=5, and min_dist=0.1. Third, HDBSCAN is employed to identify topic clusters within the reduced space. To determine the optimal number of clusters, we search for the min_cluster_size value that yields the highest topic quality, measured using the criteria proposed in Dieng et al. (2020). As shown in Fig. 9, we select min_cluster_size=250, resulting in 121 topic clusters with the best topic quality. Using these clusters, we sample abstract pairs either from within the same topic or across different topics. Table 6 summarizes the pair distribution used for the `emb2com` and `emb2dif` tasks across training, validation, and testing datasets.

```
You are an expert in biomedical
literature analysis.

You are asked to compare two PubMed
abstracts and identify their
commonalities. Please use concise
language. Please directly list five
commonalities between two abstracts.
Here are two abstracts:

1. "{abstract1}"
2. "{abstract2}"
```

Figure 7: The prompt template for listing five commonalities for a abstract pair.

```
You are an expert in biomedical
literature analysis.

You are asked to compare two PubMed
abstracts and identify their
differences. Please use concise
language. Please directly list five
differences between two abstracts.
Here are two abstracts:

1. "{abstract1}"
2. "{abstract2}"
```

Figure 8: The prompt template for listing five differences for a abstract pair.

C Training Hyperparameters & Details

For the first phase of two-phase training procedure, we freeze all model parameters except for embedding adapter $\mathcal{A}$. We then use SFTTrainer in the trl package to optimize the adapter. In the SFTConfig, we set the learning rate as 1e-3, batch size as 4, gradient accumulation steps as 8, and max sequence length as 2,048. With this settings, we train the adapter using AdamW optimizer (default parameters), linear scheduler with a warmup phase, mixed precision (i.e., bfloat16) for one epoch.

As for the second phase of two-phase training

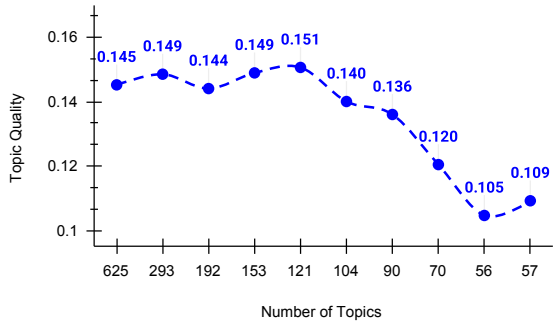


Figure 9: The line plot for helping identify the best number of topics.

Table 6: Distribution of abstract pairs used for the `emb2com` and `emb2dif` tasks across training, validation, and testing datasets. Pairs are constructed based on topic assignment using BERTopic. Each dataset contains a balanced number of same-topic and different-topic pairs to ensure diversity and control for topic-based variation.

Training Dataset		
	`emb2com`	`emb2dif`
Pairs from the Same Topic	120,897	120,897
Pairs from Different Topics	120,897	120,897
Validation Dataset		
	`emb2com`	`emb2dif`
Pairs from the Same Topic	1,562	1,562
Pairs from Different Topics	1,564	1,564
Testing Dataset		
	`emb2com`	`emb2dif`
Pairs from the Same Topic	1,589	1,589
Pairs from Different Topics	1,591	1,591

procedure and one-phase training procedure, we use SFTTrainer with LoraConfig from peft package. We set the learning rate as 5e-5, batch size as 4, gradient accumulation steps as 8, max grad norm as 1, and max sequence length as 2,048. On the other hand, we set r as 16, lora alpha as 32, lora dropout as 0.05, and bias as none. We only optimze q_proj and k_proj in M_{base} and set $\mathcal{A}$ as the module to save. We train the adapter as well as LoRA parameters using AdamW optimizer (default parameters), linear scheduler with a warmup phase, mixed precision (i.e., bfloat16) for one or two epoch(s).

We train the model with the above settings on one Nvidia H100 GPU. The training time for `ctELM` on 1.2M data using one-phase training procedure for one epoch (1P-1E) takes around 13 hours. On the other hand, the training time for `ctELM` on 1.2M data using two-phase training procedure for one epoch (2P-1E) takes around 26 hours.

Table 7: G-Eval consistency and fluency for the best-performing baseline and our model.

Model	Consistency	Fluency
Vec2Text-sect-ft	0.26±0.08	0.29±0.12
ctELM (1.2M, 5-task, 1P-1E)	**0.34±0.12**	**0.92±0.08**

D Consistency and Fluency

We measure *consistency* and *fluency* quantitatively with G-Eval (Liu et al., 2023), using the open-source DeepEval framework.[2] To define metrics, the framework requires 'criteria' and 'evaluation steps,' which we provide for consistency and fluency in Figures 10 and 11, respectively.

Table 7 shows G-Eval scores for the best-performing baseline and the `ctELM` 5-task 1P-1E model. Though consistency of both models can be improved, we find that `ctELM` has 31% higher consistency and 317% higher fluency.

To further investigate these scores, we manually review 25 outputs from Vec2Text-sect-ft and `ctELM` for errors in consistency and fluency. We analyze these qualitatively by extracting common themes and finding representative examples, as shown in Table 8.

E Automatic Plausibility Analysis

To measure plausibility of generated clinical trial abstracts under more conditions than feasible with human experts, we develop an LLM-based win rate experiment to mirror the expert version in §5.1.

E.1 Methods

As the discriminator agent, we employ `gpt-4o-2024-11-20` with the prompt shown in Fig. 12. Though an LLM discriminator may be subject to systematic self-preference (Panickssery et al., 2024), using a different model to judge than to generate may ameliorate this problem (Xu et al., 2024). To avoid the known phenomenon of positional bias in LLMS (Gu et al., 2024), we randomize whether the real abstract is first or second. Further, each win rate for each system is computed 5 times with different random seeds to ensure aggregation over different orderings of each pair.

Additionally, many clinical trial abstracts (real and generated) contain clinical trial registry identi-

[2] `https://github.com/confident-ai/deepeval`

Consistency

Criteria: "Determine whether the actual output describes the same clinical trial as the input."

Evaluation steps:

- "Check whether the medical condition in 'actual output' reflects that of 'input'."
- "Check whether the study design (e.g., randomized controlled trial, observational study) in 'actual output' reflects that of 'input'."
- "Check whether the intervention (e.g., drug, therapy) in 'actual output' reflects that of 'input'."
- "Check whether the population (e.g., age group, sex, health status) in 'actual output' reflects that of 'input'."

Figure 10: Prompts to evaluate *consistency* using G-Eval via the DeepEval framework.

Fluency

Criteria: "Determine the quality of the 'actual output' in terms of grammar, spelling, punctuation, word choice, and sentence structure."

Evaluation steps:

- "Check whether 'actual output' follows standard grammar rules"
- "Check whether 'actual output' is free of spelling and punctuation errors"
- "Check whether 'actual output' uses appropriate word choice"
- "Check whether 'actual output' has a coherent sentence structure"

Figure 11: Prompts to evaluate *fluency* using G-Eval via the DeepEval framework.

Table 8: Common types of Consistency and Fluency errors.

Consistency		
Model	**Error type**	**Examples**
ctELM	Imprecision of drugs	"Tropisetron" → "Granisetron" (both 5-HT3 antagonists used as antiemetics)
		"Telithromycin" → "Clarithromycin" (both antibiotics used to treat pneumonia)
	Incorrect patient counts	"463 patients" → "1,000 patients"
		"176 eyes of 152 patients" → "100 eyes from 50 patients"
	Simplification of multi-arm studies	Removing a third "minimal contact CB bibliotherapy" group from a study that also included (1) a 6-session Cognitive Behavioral (CB) group and (2) a control group that just got educational brochures.
Vec2Text-sect-ft	Dropping important words	"daily interruption of sedation" → "daily sedation"
	Dropping of statistical results	
	Jumbling of words and roots	"Acupuncture and needle contact were superior to control in reducing the muscle hypertonicity of all muscles except SCM"→"Muscle contact and hypertouch were superior to needle contact in reducing sclerotherapy"
	Numerical imprecision	"age=15.5 years, SD=1.2" → "mean age, 22.5+/-2.5 years"
	Hallucination of nonsensical, irrelevant phrases	"resected apnea"
		"pharmacokinetics of nisoplaban"
		"after initial cigarette-dosing"
		"a single dose of myosinophils (Meltz, n=30)"
		"cumulative femur relapse"
Fluency		
Model	**Error type**	**Examples**
ctELM	Spacing errors	"Patients'satisfaction"
	Punctuation errors	"p<001" (should have a decimal point)
Vec2Text-sect-ft	Incoherent acronyms	"low-grade tuberculosis (LO)" "early hip osteoarthritis (EE)"
	Low-complexity stretches	"S-s-s-s-s-s-s-s-s-s-s"
		"100/100/100/100/100 ml"
		"a single dose of a single dose of a single dose [...]"
	Illogical phrases	"Vitamin D deficiency is a risk factor for developing sun exposure"
	Spelling errors	"acanemia"
	Punctuation errors	"P.05" (should have a "<")

```
Which of the following abstracts is
more likely to be a real abstract
describing a clinical trial? Return
only "1" or "2".

1. "{abstract1}"

2. "{abstract2}"
```

Figure 12: The discriminator agent prompt. Order of real and generated abstracts is randomized.

fiers. As identifiers may be memorized and associated with study details in the discriminator agent's model, we redacted any such identifiers from both real and generated abstracts. We crafted regular expressions to capture each format appearing in World Health Organization's International Clinical Trials Registry Platform[3] as of February 10, 2025 (Table 9) and replaced matches with "`[redacted]`".

E.2 Results

The LLM-based win rates of abstracts from interpolated embeddings and from embeddings moved along CAVs are shown in Table 10. First, we note the overall similarity in the automatic results to human results for the conditions tested by both. Specifically, Vec2Text-sect-ft (with original embeddings) achieves a win rate of 0.02 vs. experts and 0.01 vs. the LLM, whereas `ctELM` based on `Llama-3.1-8B-Instruct` (with interpolated embeddings) achieves a win rate of 0.44 vs. experts and 0.40 vs. the LLM. LLM-based win rates are in fact slightly lower than their human counterparts, meaning they were better able to discriminate. This suggests the LLM discriminator agent is reliable enough to provide useful results for other conditions.

Second, we note that using novel embeddings has little on affect plausibility of abstracts generated by `ctELM`. For `Llama-3.1-8B-Instruct`, interpolated embeddings and those produced using age CAVs in fact have slightly higher win rates than for original embeddings. Though there are drops for interpolated vs. original embeddings for Gemma 3 models, their inter-quartile ranges still overlap. These results further suggests that `ctELM` has learned not only to create fluent text descriptions of existing clinical trials, but has learned a manifold of plausible clinical trials.

Finally, we note that win rates are lower for Gemma 3 models. This is not unexpected, given their lower Semantic Consistency. It is also in line with expectations that the large 4B parameter model performs better than the 1B parameter model. However, we see no benefit here of the continued domain-based pretraining and fine-tuning of `medgemma-4b-it`, which is suprising. It is possible that the narrow focus of this models additional training tasks (mostly multiple choice questions) affected their general language modeling or instruction following capabilities without benefiting this particular biomedical task.

F Data Collection for Sex Concept Activation Vector

To identify a symmetric axis of cohort sex, we collected clinical trials describing interventions that *could* apply to both sexes, but that happened to only include one sex in the study cohort. We thus searched PubMed for randomized controlled trials with only one of the MeSH terms 'Male' and 'Female,' and excluding studies with MeSH terms related to sex-specific conditions (such as prostate cancer) or procedures (such as hysterectomy). We further required one of four gendered nouns ('men,' 'women,' 'boys,' 'girls') to appearing in the 'Title/Abstract field' to filter out studies with single-sex MeSH terms but no mention of cohort sex in abstracts. The complete PubMed search strings are provided in Figs. 14 and 15. Search results were sorted using the 'Best Match' option, and, for each sex, the first 25 were collected that satisfied two manually verified criteria: (1) cohort sex was mentioned in the abstract (not just the title), and (2) there were no implied participants of the opposite sex, for example, "Study participants: 50 adults (23 women; 46%)." We then augmented the data to ensure semantically symmetrical pairs by using `gpt-4o-2024-11-20` (depicted as 'Augmentation agent' in Fig. 13) to reverse the sex of study participants in these initial 50 abstracts, using the prompt in Fig. 16. This created a total of 100 abstracts describing clinical trials: 25 real male, 25 real female, 25 synthetic male, and 25 synthetic female. All synthetic samples were manually reviewed for successful change of cohort sex and consistency of other study details.

[3] https://www.who.int/tools/clinical-trials-registry-platform

Regex	Registry
`ACTRN[0-9]+`	Australian New Zealand Clinical Trials Registry
`ChiCTR[A-Z0-9-]+`	Chinese Clinical Trials Register
`CTIS[0-9-]+`	European Union Clinical Trials Information System
`CTRI[0-9/]+`	Clinical Trials Registry - India
`DRKS[0-9]+`	German Clinical Trials Register
`EUCTR[0-9a-zA-Z-]+`	European Clinical Trials Register
`IRCT[0-9]+N[0-9]+`	Iranian Registry of Clinical Trials
`ISRCTN[0-9]+`	UK Clinical Study Register
`ITMCTR[0-9]+`	International Traditional Medicine Clinical Trial Registry
`JPRN-[a-zA-Z0-9]+`	Japan Primary Registries Network
`KCT[0-9]{7}`	Korean Clinical Research Information Service
`LBCTR[0-9]+`	Lebanese Clinical Trials Registry
`NCT[0-9]{8}`	US National Clinical Trial
`NL-OMON[0-9]+`	Overview of Medical Research in the Netherlands
`PACTR[0-9]+`	Pan African Clinical Trials Registry
`RBR-[a-z0-9]+`	Brazilian Clinical Trials Registry
`RPCEC[0-9]{4}`	Cuban Registry of Clinical Trials
`SLCTR/\d+/\d+`	Sri Lanka Clinical Trials Registry
`TCTR[0-9]+`	Thai Clinical Trials Registry

Table 9: Regular Expressions for identifying Clinical Trial Registry identifiers.

Table 10: Win Rate of abstracts generated from original or modified embedding vectors when presented to an LLM discriminator along with a real abstract. For `ctELM`, the 1P-1E training procedure is used with the 5-task 1.2M dataset, and all CAVs are applied with $|\alpha| = 0.5$.

		Win rate by embedding type			
Method	**Base model**	**Orig.**	**Interp.**	**CAV-sex**	**CAV-age**
Vec2Text	-	0.00±0.00	-	-	-
Vec2Text-ft	-	0.01±0.00	-	-	-
Vec2Text-sect	-	0.00±0.00	-	-	-
Vec2Text-sect-ft	-	0.01±0.00	-	-	-
ctELM	`gemma-3-1b-it`	0.12±0.02	0.13±0.05	-	-
ctELM	`gemma-3-4b-it`	0.31±0.07	0.29±0.06	-	-
ctELM	`medgemma-4b-it`	0.22±0.05	0.16±0.02	-	-
ctELM	`Llama-3.1-8B-Instruct`	**0.39±0.06**	**0.40±0.06**	**0.38±0.07**	**0.40±0.08**

G Data Collection for Age Concept Activation Vector

Similarly to the sex Concept Activation Vector, we collected clinical trials describing interventions that *could* apply to both children and aged subjects, but that happened to only include one of these groups in the trial. We thus searched PubMed for randomized controlled trials with only one of the top-level MeSH age groups 'Child' (defined in MeSH as ages 6-12) and 'Aged' (defined as age 65 and above), further excluding the other top-level groups ('Infant', 'Child, Preschool', 'Adolescent', 'Adult', and 'Middle Aged'). To find studies applicable across ages, we also excluded age-specific study elements, such as 'Schools' for child studies or 'Dementia' for aged studies. The complete PubMed search strings are provided in Figs. 17 and 18. Again, search results were sorted using the 'Best Match' option, and, for each age group, the first 25 were collected that satisfied two manually verified criteria: (1) subject age was mentioned in the abstract, and (2) there were no implied participants of other ages. We again augmented the data to ensure semantically symmetrical pairs by using `gpt-4o-2024-11-20` to reverse the age of study participants in these initial 50 abstracts, using the prompt in Fig. 19. This created a total of 100 ab-

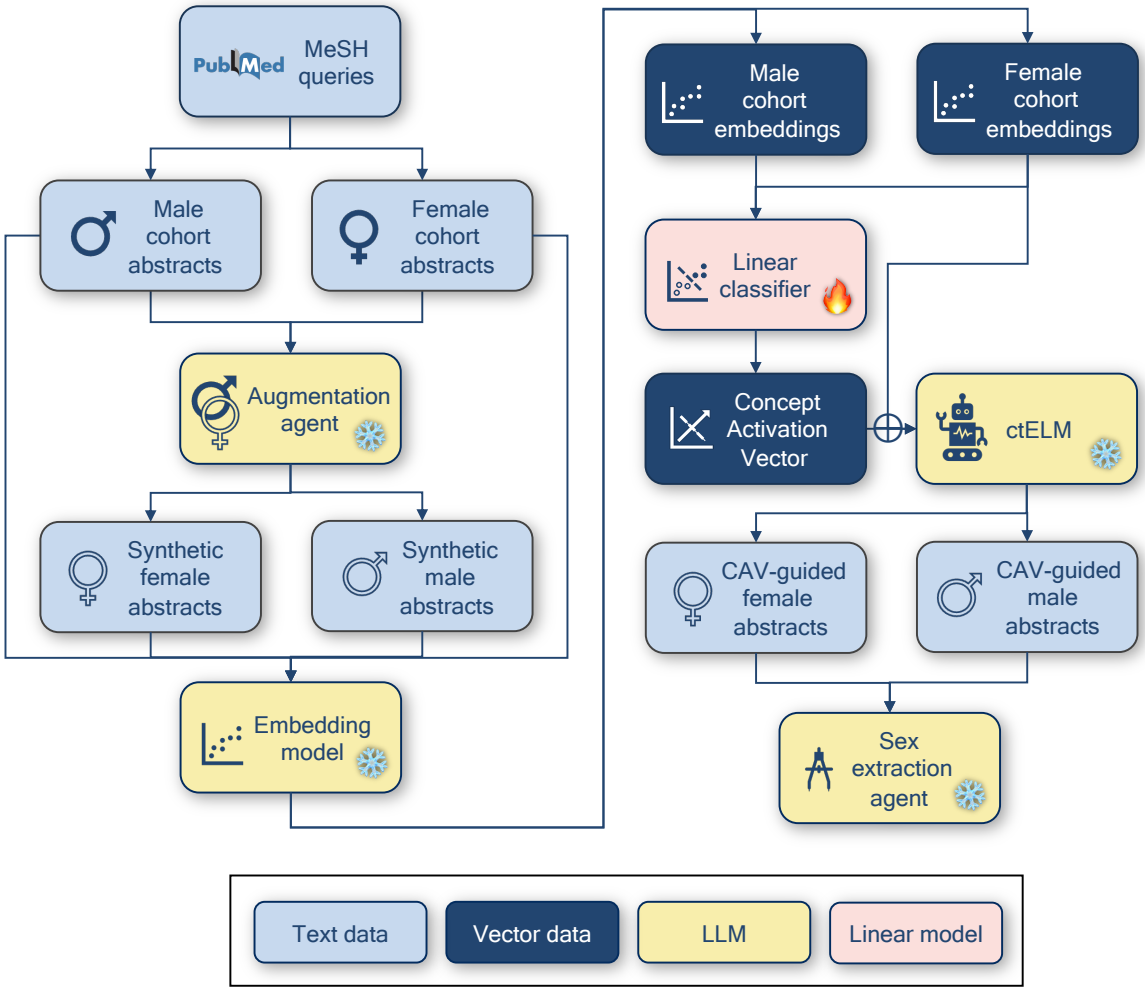


Figure 13: The workflow for modifying clinical trial embeddings with Concept Activation Vectors, including data collection, augmentation, linear model training, generation from modified embeddings, and evaluation. Depicted here is the sex CAV; the same workflow applies to age, except with child and aged instead of male and female.

stracts describing clinical trials: 25 real child, 25 real aged, 25 synthetic child, and 25 synthetic aged. All synthetic samples were manually reviewed for successful change of subject age and consistency of other study details.

H Extraction of Subject Demographics

To evaluate responsiveness of `ctELM` to CAVs, we employed an extraction agent comprising `gpt-4o-2024-11-20` with the system messages depicted in Figures 20 and 21 for sex and age, respectively. For both, the user prompt template was "Now process the following abstract: {abstract}".

I Generated Examples of Interpolated Embedding

To illustrate the generative capabilities of `ctELM` on interpolated embeddings, we present three examples of generated texts (background, objective, and result) derived from the average of two distinct abstract embeddings. As shown in Fig. 22, Fig. 23, and Fig. 24, `ctELM` successfully synthesizes semantically coherent sentences that reflect thematic overlaps between the paired source abstracts. This demonstrates the model's capacity to interpolate meaningfully between known research areas. Importantly, this capability suggests a promising avenue for exploratory scientific hypothesis generation. For instance, by sampling embeddings from underrepresented or "empty" regions of the semantic space (i.e., areas not directly covered by existing

```
(
  English[Language]
  AND (Randomized Controlled Trial [Publication Type])

  AND (Male[MeSH Terms])
  NOT (Female[MeSH Terms])

  NOT (Genitalia[MeSH Terms])
  NOT (Urogenital Diseases[MeSH Terms])
  NOT (Pelvic Neoplasms[MeSH Terms])
  NOT (Urogenital Surgical Procedures[MeSH Terms])
  NOT (Fertility Preservation[MeSH Terms])
  NOT (Contraceptive Devices[MeSH Terms])
  NOT (Alopecia[MeSH Terms])
  NOT (Gonadal Disorders[MeSH Terms])
  NOT (Gonadal Hormones[MeSH Terms])
) AND
(
  ("2024/01/01"[EPDAT] : "2024/12/31"[EPDAT])
) AND
(
  (men[Title/Abstract]) OR (boys[Title/Abstract])
)
```

Figure 14: The PubMed search string for male single-sex clinical trials.

training data), ctELM could be prompted to generate novel study hypotheses, bridging previously unconnected biomedical concepts. This highlights the potential of embedding-based generation as a tool for ideation and discovery in literature-based research.

```
(
  English[Language]
  AND (Randomized Controlled Trial[Publication Type])

  AND (Female[MeSH Terms])
  NOT (Male[MeSH Terms])

  NOT (Pregnancy[MeSH Terms])
  NOT (Menopause[MeSH Terms])
  NOT (Genitalia[MeSH Terms])
  NOT (Urogenital Diseases[MeSH Terms])
  NOT (Breast Neoplasms[MeSH Terms])
  NOT (Pelvic Neoplasms[MeSH Terms])
  NOT (Urogenital Surgical Procedures[MeSH Terms])
  NOT (Menstruation Disturbances[MeSH Terms])
  NOT (Osteoporosis, Postmenopausal[MeSH Terms])
  NOT (Fertility Preservation[MeSH Terms])
  NOT (Contraceptive Devices[MeSH Terms])
  NOT (Gonadal Disorders[MeSH Terms])
  NOT (Gonadal Hormones[MeSH Terms])
) AND
(
  ("2024/01/01"[EPDAT] : "2024/12/31"[EPDAT])
) AND
(
  (women[Title/Abstract]) OR (girls[Title/Abstract])
)
```

Figure 15: The PubMed search string for female single-sex clinical trials.

```
Modify this abstract so the subjects are {'male','female'} rather than
{'female','male'}. Output only the abstract, with no quotes or formatting.

"{abstract}"
```

Figure 16: The prompt template for symmetric augmentation of abstracts for the sex Concept Activation Vector.

```
(
  English[Language]
  AND (Randomized Controlled Trial[Publication Type])

  AND (Child[MeSH Terms])
  NOT (Child, Preschool[MeSH Terms])
  NOT (Infant[MeSH Terms])
  NOT (Adolescent[MeSH Terms])
  NOT (Adult[MeSH Terms])
  NOT (Middle Aged[MeSH Terms])
  NOT (Aged[MeSH Terms])

  NOT (Immunization Schedule[MeSH Terms])
  NOT (Child Behavior[MeSH Terms)
  NOT (Growth Disorders[MeSH Terms])
  NOT (Growth Hormone[MeSH Terms])
  NOT (Growth and Development[MeSH Terms])
  NOT (Tooth, Deciduous[MeSH Terms])
  NOT (Child Abuse[MeSH Terms])
  NOT (Family[MeSH Terms])
  NOT (Schools[MeSH Terms])
  NOT (Curriculum[MeSH Terms])
  NOT (Congenital, Hereditary, and Neonatal Diseases and Abnormalities
  [MeSH Terms])
  NOT (Neurodevelopmental Disorders[MeSH Terms])
) AND
(
  ("2024/01/01"[EPDAT] : "2024/12/31"[EPDAT])
)
```

Figure 17: The PubMed search string for child single-age-group clinical trials.

```
(
  English[Language]
  AND (Randomized Controlled Trial[Publication Type])

  AND (Aged[MeSH Terms])
  NOT (Child[MeSH Terms])
  NOT (Child, Preschool[MeSH Terms])
  NOT (Infant[MeSH Terms])
  NOT (Adolescent[MeSH Terms])
  NOT (Middle Aged[MeSH Terms])

  NOT (Breast Neoplasms[MeSH Terms])
  NOT (Dementia[MeSH Terms])
  NOT (Polypharmacy[MeSH Terms])
  NOT (Activities of Daily Living[MeSH Terms])
) AND
(
  ("2024/01/01"[EPDAT] : "2024/12/31"[EPDAT])
)
```

Figure 18: The PubMed search string for aged single-age-group clinical trials.

```
Modify this abstract so the subjects are {'children','older adults'} rather
than {'older adults','children'}. Include specific ages. Output only the
abstract, with no quotes or formatting.

"{abstract}"
```

Figure 19: The prompt template for symmetric augmentation of abstracts for the age Concept Activation Vector.

```
You are a biomedical natural language processing assistant. Given the abstract
of a clinical trial study, your task is to identify the gender of the study
population.

Your output must be in the following JSON format:
{
  "gender": "female"  // or "male" or "neutral"
}

Guidelines:
- If the abstract mentions that the study participants are women or females,
output "female".
- If the abstract mentions men or males, output "male".
- If the abstract only mentions the number of participants without specifying
gender, output "neutral".
- If both male and female participants are mentioned and the study includes
both, still output "neutral".
- Do not infer gender based on disease or context. Only use explicit statements.
```

Figure 20: The subject sex extraction agent system message.

```
You are a biomedical natural language processing assistant. Given the abstract
of a clinical trial study, your task is to extract or infer the average age (in
years) of the study population.

Your output must be in the following JSON format:
{
  "age": 54.3  // numerical value only
}

Guidelines:
1. If the study mentions the **mean or average age**, extract and return that
value.
2. If the study mentions an **age range** (e.g., "30 to 50 years"), compute the
average (e.g., (30+50)/2 = 40.0) and return that value.
3. If no explicit age value is mentioned, infer the most likely average age
based on population group terms in the text, using this mapping:

- "Child, Preschool": 2-5 years  -> 3.5
- "Child": 6-12 years  -> 9
- "Adolescent": 13-18 years  -> 15.5
- "Adult": 19-44 years  -> 31.5
- "Middle Aged": 45-64 years  -> 54.5
- "Aged": 65+ years  -> 75
- "Aged, 80 and over": 80+ years  -> 85
- "Octogenarians": 80-89 years  -> 84.5
- "Nonagenarians": 90-99 years  -> 94.5
- "Centenarians": 100+ years  -> 100

Choose the most appropriate inferred value if only a population label is
present.

Only include the JSON output. Do not explain or add commentary.
```

Figure 21: The subject age extraction agent system message.

Example 1: Generated Background of Interpolated Embedding

Generated Background Section of Interpolated Embedding between PMID=24099432 and PMID=17064200: "This paper presents baseline data from a randomized clinical trial examining the effectiveness of a *cognitive behavioral intervention (CBI)* for improving medication adherence and depression outcomes among *patients with poorly controlled hypertension*."

Partial Abstract for PMID=24099432 (Cognitive Behavioral Therapy, Depressive disorder): "[Objective] We tested whether a brief cognitive behavioral (CB) group and bibliotherapy prevention reduce major depressive disorder (MDD) onset, depressive symptoms, and secondary outcomes relative to brochure controls in adolescents with self-reported depressive symptoms when school personnel recruit participants and deliver the intervention. ... [Results] The finding that a brief CB group intervention delivered by real-world providers significantly reduced MDD onset relative to both brochure control and bibliotherapy is very encouraging, although effects on continuous outcome measures were small or nonsignificant and approximately half the magnitude of those found in efficacy research, potentially because the present sample reported lower initial depression."

Partial Abstract for PMID=17064200 (Hypertension): "[Objective] To examine potential threats to internal and external study validity caused by differential patient withdrawal from a randomized controlled trial evaluating pharmacist management of hypertension, to compare the characteristics of patients who withdrew with those of patients who completed the study, and to identify characteristics that predispose patients to withdraw from hypertension management. ... [Results] Therefore, internal validity was preserved, and outcomes from the study groups could be reliably compared. A lack of significant differences between patients who withdrew versus those who completed, with the exception of insurance status, suggests that external validity was not jeopardized."

Figure 22: Example of generated background from interpolated embedding between two clinical trials. The generated text reflects a synthesis of themes related to cognitive behavioral therapy and hypertension.

Example 2: Generated Objective Section of Interpolated Embedding

Generated Objective of Interpolated Embedding between PMID=15914575 and PMID=9777179: "A study was conducted to determine if irritable bowel syndrome (IBS) patients with depressive symptoms have a better response to a selective serotonin reuptake inhibitor (SSRI) than those without depressive symptoms."

Partial Abstract for PMID=15914575 (Irritable bowel syndrome): "[Background] Melatonin, a sleep promoting agent, is involved in the regulation of gastrointestinal motility and sensation. [Objective] We aimed to determine if melatonin was effective in improving bowel symptoms and sleep disturbances in irritable bowel syndrome (IBS) patients with sleep disturbance. ... [Results] The findings suggest that the beneficial effects of melatonin on abdominal pain in IBS patients with sleep disturbances are independent of its action on sleep disturbances or psychological profiles."

Partial Abstract for PMID=9777179 (Chronic mental illness): "[Objective] We sought to identify baseline predictors of response to clozapine. Data were from a 15-site randomized clinical trial comparing clozapine and haloperidol in hospitalized patients with refractory schizophrenia (n = 423). Three-month outcomes were analyzed with the full sample (n=368 due to attrition). ... [Results] Although high levels of symptoms were associated with greater improvement on clozapine, these findings are not robust enough to suggest that any specific, clinically defined subgroup of refractory patients should be preferentially targeted for clozapine treatment."

Figure 23: Generated objective from the interpolation of abstract embeddings concerning irritable bowel syndrome and chronic mental illness. The output proposes a novel trial scenario integrating elements of both parent studies.

Example 3: Generated Results Section of Interpolated Embedding

Generated Result of Interpolated Embedding between PMID=15166570 and PMID=12860586: "The results of this study demonstrate that patients with functional constipation have a better outcome after treatment with a single dose of subcutaneous diamorphine than after treatment with a single dose of subcutaneous hyoscine butylbromide."

Partial Abstract for PMID=15166570 (Opioids, Morphine, Urinary tract dysfunction): "[Background] Intrathecal administration of opioids may cause lower urinary tract dysfunction. In this study, the authors compared the effects of morphine and sufentanil administered intrathecally in a randomized double-blind fashion (two doses each) on lower urinary tract function in healthy male volunteers. ... [Conclusion] Intrathecal opioids decrease bladder function by causing dose-dependent suppression of detrusor contractility and decreased sensation of urge. Recovery of normal lower urinary tract function is significantly faster after intrathecal sufentanil than after morphine, and the recovery time is clearly dose dependent."

Partial Abstract for PMID=12860586 (Dyspepsia): "[Background] The value of the test-and-treat strategy in the approach to dyspepsia has been evaluated only in a few secondary care studies. Most patients with dyspepsia, however, are treated by their primary care physician ... [Conclusion] The test-and-treat strategy proved to be as effective and safe as prompt endoscopy. Only a minority of patients were referred for endoscopy after the test-and-treat approach."

Figure 24: Example of a generated result sentence from interpolated embeddings of abstracts on opioids and dyspepsia. The output blends insights into drug response and gastrointestinal outcomes, demonstrating semantic consistency across domains.

Uncertainty-Aware Multi-Label Routing of Clinical Text to Surveillance Pathways

Agathe Zecevic[1,2], Angus Roberts[3], Sebastian S. Zeki[1,4]
[1]Faculty of Life Sciences & Medicine, King's College London, United Kingdom
[2]Clinical Scientific Computing, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
[3]Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom
[4]Gastroenterology Department, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
agathe.zecevic2@nhs.net

Abstract

Clinical decision support systems that operate across multiple downstream care pathways must first determine which pathway or pathways are relevant for a given patient. We study this routing problem in gastrointestinal surveillance, where paired endoscopy and histopathology text reports may indicate multiple concurrent conditions and therefore require multi-label routing. In this context, standard hard-label evaluation can be insufficient: a model may achieve reasonable overall performance while still excluding clinically important pathways when uncertain.

We formulate gastrointestinal report routing as a multi-label uncertainty-aware classification task over six pathway labels and compare lightweight lexical baselines, frozen embedding models and a fine-tuned transformer baseline under two complementary uncertainty mechanisms: threshold-based abstention and set-valued conformal prediction. Using 1,773 paired reports from a single NHS trust with disjoint train, calibration and test splits, we evaluate both hard-routing performance and the downstream review burden introduced by uncertainty-aware prediction.

The fine-tuned ClinicalBERT model achieved the strongest overall performance (0.811 subset accuracy, 0.861 macro-F1) and the lowest AURC of 0.084 under min-margin abstention. Threshold-based abstention consistently reduced exact-match routing error on accepted reports. For conformal routing at $\alpha = 0.10$, Mondrian calibration achieved high mean positive-label recall coverage across learned baselines (0.883-0.917). The fine-tuned model achieved 0.891 mean recall coverage with a mean prediction set size of 1.70, 0.642 candidate-label precision and 0.61 false-positive labels per report. Compared with a recall-tuned threshold baseline at similar recall, Mondrian CP produced smaller candidate sets, higher candidate-label precision and fewer false-positive pathway suggestions.

These results show that uncertainty-aware evaluation exposes clinically important failure modes missed by aggregate metrics. They also show that high-recall routing is not cost-free: set-valued prediction can reduce missed-pathway risk but must be interpreted as candidate generation for downstream review rather than automated pathway selection.

1 Introduction

Text-based clinical decision support systems use information recorded in free-text clinical documents to assist with screening, surveillance, triage and follow-up decisions (Sutton et al., 2020). In cases where a single system supports multiple possible downstream actions, an initial routing step is often required to determine which clinical pathway or pathways are relevant for a given patient. We refer to this as the task of *routing clinical text to clinical pathways*.

This task exposes an evaluation gap between NLP-in-the-lab and NLP-in-deployment. Standard multi-label models are typically assessed using aggregate point-prediction metrics such as accuracy or F1-score. However, point prediction alone can be insufficient in health-related use-cases: a model may achieve reasonable overall performance while still excluding a clinically important pathway on the cases where it is most uncertain. A hard routing decision returns a fixed label set, whereas a set-valued prediction can instead return a small candidate set of plausible pathways. An abstaining system can defer highly uncertain cases for manual review. These mechanisms address a question that aggregate point-prediction metrics do not fully capture: not only *what* the model predicts on average, but also *how reliably* it behaves when uncertainty is clinically important. This motivates uncertainty-aware evaluation frameworks for clinical NLP, particularly for multi-label tasks in which several pathways may be simultaneously relevant.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 181–190
July 3-4, 2026 ©2026 Association for Computational Linguistics

We study this problem in gastrointestinal (GI) surveillance. Clinical decision support systems for GI surveillance (Zecevic et al., 2024; Peterson et al., 2021; Song et al., 2022) increasingly target individual conditions and ground themselves in national guidance such as Barrett's oesophagus, colorectal polyps (Fitzgerald et al., 2014; Rutter et al., 2020), or gastric atrophy (Banks et al., 2019). These conditions are clinically important because they are associated with increased risk of disease progression or gastrointestinal malignancy and therefore require pathway-specific follow-up. Decisions often depend on combining information from endoscopy reports describing macroscopic procedure findings, with histopathology reports providing microscopic confirmation from biopsies taken at procedure time. In practice, a single endoscopy-histopathology report pair can indicate multiple concurrent conditions. For instance, incidental colorectal polyps can be discovered during IBD surveillance, making routing inherently a multi-label problem. Misrouting carries significant clinical consequences: a Barrett's patient erroneously sent to the *Normal* pathway may miss surveillance intervals, whereas routing an incomplete or abandoned procedure to an inappropriate Gastric Atrophy pathway wastes downstream resources (Codipilly et al., 2018; White and Banks, 2022).

We therefore formulate GI report routing as a multi-label classification task with two complementary uncertainty-handling mechanisms. First, we study *abstention*, where the model defers cases that are too uncertain for automated routing. Second, we study *set-valued conformal prediction*, where the model returns a small candidate set of pathways designed to better retain clinically relevant labels under uncertainty. We make three contributions:

- We formulate GI surveillance pathway routing from paired endoscopy-histopathology reports as a multi-label uncertainty-aware classification problem.
- We show that threshold-based abstention substantially reduces exact-match routing error on accepted cases across lightweight and transformer-based baselines.
- We show that positive-only Mondrian conformal calibration provides a high-recall candidate-generation layer for clinically important pathways, while explicitly quantifying the associated false-positive review burden.

2 Related Work

Clinical text routing and triage. Prior work on clinical text routing and triage spans several adjacent settings. NLP has been used in referral workflows to predict target speciality from referral notes and triage referrals (Spasic and Button, 2020). Free-text triage has also been widely studied in emergency care (Stewart et al., 2023). However, these studies generally route a single text stream toward one operational endpoint rather than assigning paired reports to potentially multiple concurrent pathways.

Selective prediction in NLP. Selective prediction adds an abstention layer to traditional classifiers, producing risk-coverage trade-offs (El-Yaniv and Wiener, 2010). Wen et al. (2025) survey selective prediction in the context of LLMs, summarising common approaches such as threshold-based and learned selection functions. Fisch et al. (2022) address multi-label settings with conformal guarantees that limit false-positive predictions, which is critical for domains like healthcare.

Conformal prediction. Split conformal prediction provides distribution-free marginal coverage guarantees by calibrating nonconformity scores on a held-out set (Vladimir Vovk, 2005; Angelopoulos and Bates, 2022). The Mondrian extension refines this idea by yielding label-conditional guarantees that are especially relevant under class imbalance or asymmetric misrouting costs. Campos et al. (2024) survey conformal prediction for NLP, including both marginal and class-conditional variants. In clinical applications, conformal methods have recently been explored for medical multiple-choice question answering (Ke et al., 2025), medical text mining for epidemiological surveillance (Genari and Goedert, 2025), multi-label ICD coding from clinical narratives (Zhang et al., 2025a) and medical entity extraction from radiology text (Shrestha and Kim, 2026).

This study is distinct from our prior Barrett's oesophagus decision-support work, which focused on condition-specific surveillance scheduling. This work studies the upstream multi-label routing problem across six GI pathways using a newly annotated routing dataset.

3 Methods

3.1 Task Formulation

Given a merged endoscopy-histopathology report x, we predict a label set $\hat{Y} \subseteq \mathcal{L}$

where $\mathcal{L} = \{$*Normal*, *Barrett's*, *Colorectal Polyps*, *Gastric Atrophy*, *Other*, *Incomplete*$\}$. A selective predictor (f, g) consists of a base classifier f and a selection function $g : \mathcal{X} \to \{0, 1\}$ that determines whether to accept ($g(x) = 1$) or abstain ($g(x) = 0$). We decompose routing into six independent binary decisions, one per label, with associated probability scores $f_j(x) \in [0, 1]$. We evaluate via risk-coverage curves: *selective risk*, which is defined as the error rate on accepted instances versus *coverage*, defined as the fraction of instances accepted. For the point-prediction baselines, each label j is predicted positive when $f_j(x) \geq 0.5$.

Label logic. Gold labels are multi-label, but not all label combinations are clinically coherent. In our case, the label *Normal* denotes the absence of any pathway-relevant clinical finding and therefore does not co-occur with any other label in the annotated datasets.

3.2 Baselines

We compare four clinical pathway routing baselines with increased computational complexity.

B0: Keyword baseline This pipeline consists of regex patterns for each class matched against the merged report text. We incorporate negation cues (such as "no evidence of", "history of", "to rule out") to improve baseline performance. We derive heuristic scores for each label as the number of matched patterns divided by the total number of patterns defined for that label.

B1: TF-IDF + Logistic Regression. We split each report into affirmed and negated text spans using the same negation cues as B0 and fit separate TF-IDF vectorisers to each channel. For each label, the regularisation strength $C \in \{0.01, 0.1, 1.0, 10.0\}$ was selected by 3-fold cross-validation on the training split only, using macro-F1 as the selection criterion. Per-label probability calibration was performed on the training split only via Platt scaling with stratified K-fold cross-validation (K = 3). The positive and negative channels use feature budgets of 10,000 and 5,000 features respectively and are concatenated into a 15,000-dimensional representation. Both vectorisers use n-grams of order 1-2.

B2: Frozen BGE-Large embeddings + Linear head. The frozen-embedding baseline extracts fixed representations from BGE-Large-EN-v1.5, a text encoder pre-trained with contrastive learning (Xiao et al., 2024). Each report is tokenised to a maximum of 512 subword tokens and encoded with the fixed model weights. A masked mean pool over the last hidden states of non-padding tokens produces a single 1024-dimensional vector per report. A separate one-vs-rest logistic regression classifier is then fitted per label on these frozen embeddings.

B3: Fine-tuned ClinicalBERT. We fine-tune ClinicalBERT (Alsentzer et al., 2019) for multi-label classification by adding a linear classification head on top of a masked mean-pool representation of the encoder's last hidden state. The head produces six independent logits, trained with binary cross-entropy. We optimise with AdamW (weight decay 0.01) using a linear warmup over the first 10% of training steps followed by linear decay to zero. Gradients are clipped to a maximum norm of 1.0. Learning rate and number of epochs are selected jointly by grid search over $10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}$ and up to 15 epochs, evaluated by macro F1 on a 15% hold-out subset of the training split kept strictly separate from the calibration data.

For each baseline, we compute results at a uniform threshold of 0.5 and at per-label thresholds selected on the calibration set to achieve at least 90% recall for each label independently.

3.3 Selective Prediction

Threshold abstention. We use a *min-margin* selection function: accept report x if

$$g_\tau(x) = \min_j \max\big(f_j(x),\, 1 - f_j(x)\big) \;\geq\; \tau, \quad (1)$$

where $\tau \in [0.5, 1]$. This score equals the confidence of the least certain label: it falls to 0.5 whenever any label score is close to the decision boundary, so a report is abstained as soon as any single label is uncertain. This aligns the selection function directly with exact-match risk, which fails whenever any label is wrong. By contrast, max-confidence ($\max_j f_j$) can accept a case with one highly confident label while several others remain uncertain, which is an undesirable behaviour in our setting. We vary τ to generate risk-coverage curves and report the Area Under the Risk-Coverage curve (AURC) (Geifman et al., 2019), where a lower AURC indicates better selective classification.

3.4 Conformal Prediction

Conformal prediction (Vladimir Vovk, 2005) is a distribution-free framework for constructing prediction sets with finite-sample coverage guarantees, requiring only data exchangeability. We apply it

independently per label j and evaluate whether the returned candidate set retains truly present labels. For each label, the main quantity of interest is the fraction of positive test examples for which label j is included in the prediction set. We report this as per-label recall coverage:

$$P\big(j \in \mathcal{C}(x) \mid Y_j = 1\big) \;\geq\; 1 - \alpha, \qquad (2)$$

where $\mathcal{C}(x) \subseteq \mathcal{L}$ is the prediction set returned for report x. We report empirical positive-class inclusion as the fraction of positive test examples for which label *j* is included in the prediction set, we do not make any joint coverage claim across labels or across the full multi-label set.

Split conformal prediction. We apply split conformal prediction independently per label j, using the LAC (Least Ambiguous Classifier) nonconformity score as implemented in MAPIE (Taquet et al., 2022):

$$s_j(x, y_j) = 1 - f(x)_{j,y_j}, \qquad (3)$$

where $f(x)_{j,y_j}$ is the model's predicted probability for the *true* class $y_j \in \{0, 1\}$ of label j. For each label j, we set the calibration threshold $\hat{q}_j$ to the $\lceil (n_j + 1)(1 - \alpha) \rceil / n_j$ empirical quantile of the calibration scores $\{s_j(x_i, y_{j,i})\}_{i=1}^{n_j}$, computed either over all calibration examples (marginal calibration) or over the positive calibration examples only (Mondrian calibration). Label j is included in the prediction set if

$$f_j(x) \geq 1 - \hat{q}_j, \qquad (4)$$

that is, if the model assigns sufficiently high probability to the positive class.

Marginal versus Mondrian calibration. Under *marginal* calibration, $\hat{q}_j$ is estimated from all calibration examples for label j, mixing positive and negative instances. This does not directly target the conditional quantity in Eq. (2); under class imbalance, it can therefore yield poor positive-class inclusion for rare labels. For a minority label such as GA in our case, most calibration examples are negative and the model typically assigns them small positive-class probabilities, which leads to calibration scores clustering near 0. In our experiments, this noticeably reduced GA inclusion under marginal calibration at $\alpha = 0.10$. This is mitigated in *Mondrian* calibration, where $\hat{q}_j$ is instead estimated only from the positive calibration examples for label j, matching the positive-class inclusion objective in Eq. (2).

Label	Train ($n = 789$)		Cal. ($n = 523$)		Test ($n = 461$)	
	n	%	*n*	%	*n*	%
Normal	77	9.8	53	10.1	45	9.8
Barr	185	23.4	65	12.4	50	10.8
CP	115	14.6	53	10.1	53	11.5
GA	63	8.0	30	5.7	42	9.1
Other	500	63.4	361	69.0	314	68.1
Incomplete	103	13.1	50	9.6	60	13.0
Multi-label (≥ 2)	249	31.6	89	17.0	101	21.9
Mean label cardinality	1.32		1.17		1.22	

Table 1: Per-label prevalence across datasets. *Multi-label* reports cases with two or more labels. *Mean label cardinality* is the average number of labels per report.

Abstention rule. We add a separate deployment heuristic on top of conformal prediction. A report is abstained when its prediction set size $|\mathcal{C}(x)| > K$, with $K = 3$. This size threshold is not part of the conformal guarantee but a pragmatic rule intended to defer highly ambiguous cases to manual review. We chose $K = 3$ a priori based on the observed degree of clinically plausible co-occurrence in the training and calibration sets. We evaluate at $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ and report per-label empirical recall coverage (Eq. (2)), mean prediction set size and abstention rate.

4 Experiments

4.1 Data

We use six months of anonymised paired endoscopy and histopathology reports from an NHS hospital trust, merged into single documents per procedure. Data acquisition was authorised through institutional board review (IRAS ID = 257283). We remove patients who had opted out of their data being used (NHS Data Opt-Out). We construct three non-overlapping splits: a training set ($n = 789$), a calibration set ($n = 523$) and a test set ($n = 461$). The training set only is enriched for the minority classes (Barr, GA, Incomplete), the calibration and test set are not enriched.

Reports were manually annotated by a clinical researcher using labels from $\mathcal{L}$. Ambiguous cases (for instance "history of Barrett's" with no active findings) were resolved via predefined annotation rules. Table 1 reports the label distribution across datasets and the average number of labels per merged report. More details about the annotation rules and label co-occurrence can be found in Appendix A.

4.2 Results

Baseline comparison. Table 2 compares the four routing baselines on exact-match subset accuracy,

Model	Subset Acc. [95% CI]	$F1_{macro}$ [95% CI]	$Precision_{macro}$ [95% CI]	AURC [95% CI]
Keyword	0.538 [0.492, 0.584]	0.705 [0.670, 0.737]	0.646 [0.612, 0.682]	0.3718 [0.3183, 0.4264]
TF-IDF	0.646 [0.605, 0.692]	0.782 [0.743, 0.817]	0.840 [0.799, 0.880]	0.1683 [0.1318, 0.2119]
Embedding	0.705 [0.666, 0.746]	0.776 [0.733, 0.814]	**0.898** [0.861, 0.932]	0.1338 [0.1026, 0.1663]
Fine-tuned	**0.811** [0.777, 0.846]	**0.861** [0.829, 0.891]	0.873 [0.837, 0.910]	**0.0836** [0.0578, 0.1127]

Table 2: Baseline comparison on point-prediction and selective prediction performance. Reported metrics are subset accuracy, macro F1, macro precision and area under the risk-coverage curve (AURC). 95% confidence intervals were computed by non-parametric bootstrap over test reports with 1,000 resamples.

macro F1, and AURC. TF-IDF is a competitive lightweight baseline, achieving 0.646 subset accuracy and 0.782 macro-F1. The frozen BGE-Large embedding model improves subset accuracy to 0.705 and exhibits the highest macro precision (0.898). The fine-tuned ClinicalBERT model performs best overall, reaching 0.811 subset accuracy, 0.861 macro-F1 and the lowest AURC (0.0836). Paired bootstrap comparisons in Appendix B support the improvement of the fine-tuned model over the other learned baselines.

Selective prediction. Figure 1 shows risk–coverage curves under min-margin abstention. All models show decreasing selective risk as coverage decreases. This indicates that the confidence score is informative about cases where hard routing is more reliable. The fine-tuned model has the lowest AURC (0.084), followed by the frozen embedding model (0.134) and TF-IDF (0.168). The keyword baseline remains substantially weaker both at full coverage and under abstention, which can be explained by how its confidence scores are computed.

Conformal prediction. Table 3 summarises Mondrian conformal routing at $\alpha = 0.10$ for the learned baselines. Mean recall coverage remains high across models, ranging from 0.883 for the frozen embedding model to 0.917 for TF-IDF. The operational burden differs across baselines. The fine-tuned model provides the best trade-off, with the highest set precision (0.642), the smallest mean set size (1.70), the lowest false-positive burden (0.61 FP labels/report) and no abstentions under the $K = 3$ rule. These results show why recall coverage alone is insufficient: high-recall conformal routing must be interpreted together with set precision and false-positive pathway burden.

Figure 2 shows the conformal operating trade-off. At the set level (Panel A), we show that decreasing α increases mean recall coverage but reduces candidate-label precision. The learned models follow similar trade-off curves, with the fine-tuned model providing the best operating point at $\alpha = 0.10$ in Table 3. Panel B uses GA to illustrate the clinical cost of high-recall routing: Mondrian CP substantially increases GA recall relative to the hard classifier but with a marked reduction in precision.

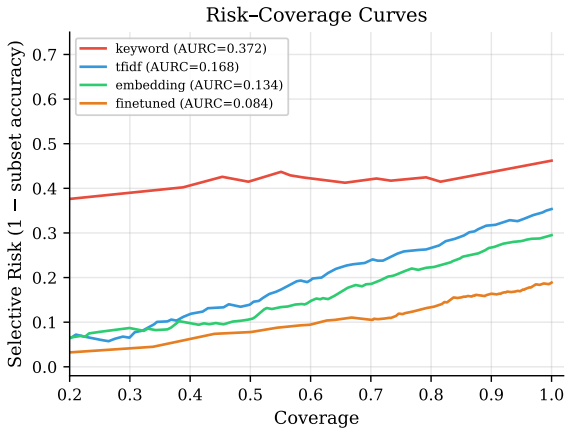


Figure 1: Selective risk-coverage curves under min-margin abstention. The low-coverage region (<0.2) is omitted as estimates become unstable with few accepted examples. AURC is computed over the full coverage range.

Model	Recall cov. ↑	Set prec. ↑	Mean set ↓	FP/report ↓	Abstain % ↓
TF-IDF	**0.917**	0.589	1.89	0.77	1.3
Embedding	0.883	0.607	1.79	0.70	1.1
Fine-tuned	0.891	**0.642**	**1.70**	**0.61**	**0.0**

Table 3: Operational summary of Mondrian conformal routing at $\alpha = 0.10$. Recall cov. is mean per-label positive-class inclusion coverage over test positives. Set precision is the candidate-label precision, defined as $\sum_i |\mathcal{C}_i \cap Y_i| / \sum_i |\mathcal{C}_i|$. FP/report is the average number of false-positive pathway labels included per report.

Per-class analysis. Table 4 compares hard prediction, a recall-tuned threshold baseline, marginal CP and Mondrian CP for the fine-tuned model at $\alpha = 0.10$. The recall-tuned threshold baseline achieves high mean recall (0.895), but candidate-label precision falls to 0.540, mean set size increases to 2.03 and the false-positive burden rises to 0.93 labels/report. Mondrian CP achieves a similar mean recall (0.891) with higher candidate-label precision (0.642), smaller sets (1.70) and fewer false-positive labels per report (0.61).

The GA row shows the main minority-label trade-off. Hard prediction has relatively high GA precision (0.80) but lower recall (0.67). Mondrian

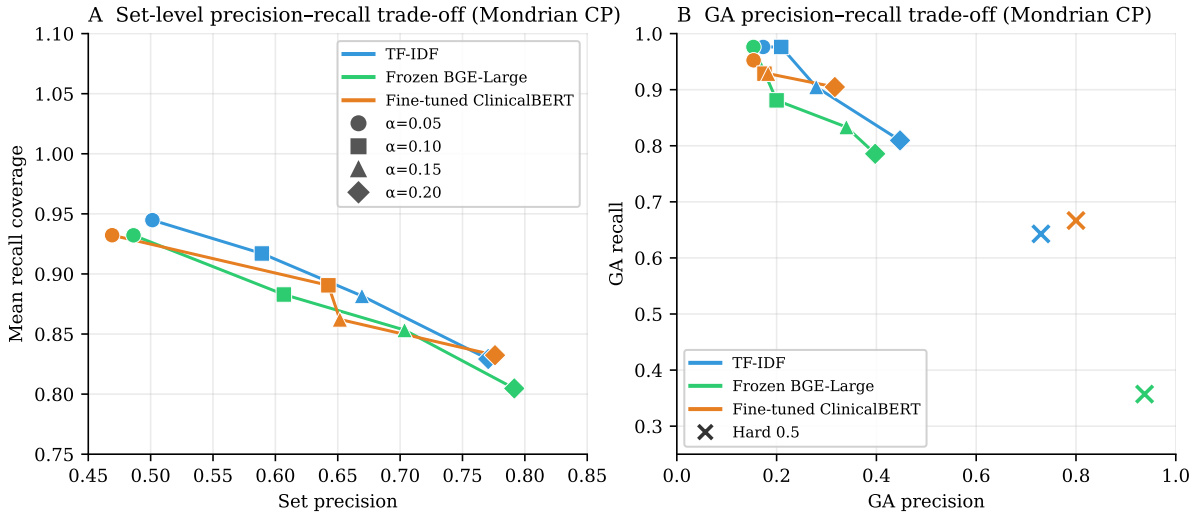


Figure 2: Precision-recall trade-offs for Mondrian conformal routing. **A:** Set-level trade-off across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ for the learned baselines, where precision is candidate-label precision and recall is mean per-label positive-class recall coverage. **B:** GA-specific precision–recall trade-off under Mondrian CP. The hard 0.5-threshold classifiers are shown as reference points.

CP increases GA recall to 0.93, but GA precision falls to 0.18, indicating that the conformal layer acts as a high-recall safety net rather than a final automatic GA decision. Marginal CP under-recovers positives for several labels. These results support interpreting conformal routing as candidate generation for downstream review.

5 Discussion and Conclusion

We evaluated a range of routing baselines with different computational requirements, motivated by the practical constraint that not all hospitals can deploy the same class of model. Even relatively lightweight approaches such as TF-IDF achieved reasonable overall performance on our annotated dataset. We highlight the fact that standard point-prediction metrics alone are not sufficient for this task as they do not characterise the risk of excluding clinically relevant pathways from a set-valued prediction. Some misclassifications have important clinical consequences, particularly false negatives on pathways where missing the correct route could prevent the relevant surveillance logic from being applied at all.

This motivated two separate uncertainty-handling strategies at inference time. First, we studied threshold-based abstention to reduce hard-routing errors on accepted cases. Second, we studied conformal set prediction to reduce the risk that clinically important pathways are excluded from the returned candidate set. Recall on clinically important pathways is more crucial than forcing a single-label decision in every case. Returning a small set of plausible pathways may be preferable to excluding the true one, because an irrelevant downstream pathway may simply yield no action, whereas omission of the correct pathway risks missing the appropriate follow-up entirely.

Our results support three conclusions. First, threshold-based abstention reduces exact-match routing error on accepted cases, with the strongest selective performance obtained by the fine-tuned model. Second, set-valued conformal prediction changes the evaluation question from whether a single hard label set is correct to whether clinically important pathways are retained for review. Third, high-recall routing is not cost-free. While Mondrian CP can recover minority pathways such as GA, this comes at the cost of increasing false-positives and therefore downstream review burden.

This distinction is important for deployment. A conformal set should not be interpreted as a final diagnosis or automatic pathway activation. Instead, it is a candidate-generation layer: labels included in $\mathcal{C}(x)$ are pathways that should not be safely excluded without further pathway-specific logic or manual review. Under this interpretation, Mondrian CP provides a useful safety mechanism because it reduces missed-pathway risk while making the review burden explicit. The comparison with a recall-tuned threshold baseline further showed that in our case, Mondrian CP achieved a better recall-

Method	Overall operating point					Per-label recall / precision					
	Mean recall	Set precision	Mean set	FP/report	Abstain %	Normal	Barrett's	CP	GA	Other	Incomplete
Hard 0.5	0.852	0.903	1.20	0.12	—	0.91 / 0.84	0.78 / 0.87	0.96 / 0.89	0.67 / 0.80	0.92 / 0.93	0.87 / 0.91
Tuned on cal. (≥90% recall)	0.895	0.540	2.03	0.93	—	0.93 / 0.40	0.74 / 0.86	0.96 / 0.80	0.95 / 0.11	0.90 / 0.93	0.88 / 0.88
Marginal CP	0.726	0.894	1.15	0.12	0.0	0.80 / 0.88	0.54 / 0.84	0.85 / 1.00	0.43 / 1.00	0.96 / 0.87	0.78 / 0.96
Mondrian CP	0.891	0.642	1.70	0.61	0.0	0.91 / 0.44	0.74 / 0.86	0.96 / 0.80	0.93 / 0.18	0.90 / 0.93	0.90 / 0.89

Table 4: Operating-point and per-label performance for the fine-tuned model at $\alpha = 0.10$

burden trade-off than simple recall-thresholding for the fine-tuned model.

Limitations. Our study uses reports from a single NHS site, in English, with a six-class routing schema, labelled by a single annotator. One class (*Other*) is heavily represented as it is a deliberate catch-all for conditions outside the current surveillance scope (for instance IBD, strictures, malignancy). The calibration set remains small for the rarest classes. With few positive calibration examples per minority class, the Mondrian threshold $\hat{q}_j$ can only take a small number of distinct values, making coverage insensitive to small changes in α.

Future work. We plan to add a local LLM zero-shot/few-shot baseline expanding on already existing work for single conditions (Zhang et al., 2025b), extend the routing schema to additional clinical pathways currently labelled as *Other* and validate the system prospectively on several temporally distant test sets.

Acknowledgments

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King's College London, with additional support from the South East London Cancer Alliance (SELCA).

This research was supported by the Clinical Research Data and Analytics platform within Research and Development, Guy's and St Thomas' NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS.

Code and Data availability

The datasets generated or analysed, or both during this study are not publicly available owing to ethical restrictions. Guy's and St Thomas' Electronic Research Records Interface (GERRI) is governed by the overarching Clinical Research Analytics Governance Group (CRAG).

The code used to run the experiments and reproduce the analyses can be made available upon reasonable request.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.

Matthew Banks, David Graham, Marnix Jansen, Takuji Gotoda, Sergio Coda, Massimiliano di Pietro, Noriya Uedo, Pradeep Bhandari, D Mark Pritchard, Ernst J Kuipers, Manuel Rodriguez-Justo, Marco R Novelli, Krish Ragunath, Neil Shepherd, and Mario Dinis-Ribeiro. 2019. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut*, 68(9):1545–1575.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal Prediction for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Don Chamil Codipilly, Apoorva Krishna Chandar, Siddharth Singh, Sachin Wani, Nicholas J. Shaheen, John M. Inadomi, Amitabh Chak, and Prasad G. Iyer. 2018. The Effect of Endoscopic Surveillance in Patients With Barrett's Esophagus: A Systematic Review and Meta-analysis. *Gastroenterology*, 154(8):2068–2086.

Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2022. Conformal Prediction Sets with Limited False Positives.

Rebecca C Fitzgerald, Massimiliano di Pietro, Krish Ragunath, Yeng Ang, Jin-Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V Kaye, Scott Sanders, Maria O'Donovan, Elizabeth Bird-Lieberman, Pradeep Bhandari, Janusz A Jankowski, Stephen Attwood, Simon L Parsons, Duncan Loft, Jesper Lagergren, Paul Moayyedi, and 2 others. 2014. British Society of Gastroenterology guidelines on the

diagnosis and management of Barrett's oesophagus. *Gut*, 63(1):7–42.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers.

Juliano Genari and Guilherme Tegoni Goedert. 2025. Mining Unstructured Medical Texts With Conformal Active Learning.

Yusong Ke, Hongru Lin, Yuting Ruan, Junya Tang, and Li Li. 2025. Correctness Coverage Evaluation for Medical Multiple-Choice Question Answering Based on the Enhanced Conformal Prediction Framework.

Emma Peterson, Folasade P. May, Odet Kachikian, Camille Soroudi, Bita Naini, Yuna Kang, Anthony Myint, Gordon Guyant, Joann Elmore, Roshan Bastani, Cleo Maehara, and William Hsu. 2021. Automated identification and assignment of colonoscopy surveillance recommendations for individuals with colorectal polyps. *Gastrointestinal Endoscopy*, 94(5):978–987.

Matthew D Rutter, James East, Colin J Rees, Neil Cripps, James Docherty, Sunil Dolwani, Philip V Kaye, Kevin J Monahan, Marco R Novelli, Andrew Plumb, Brian P Saunders, Siwan Thomas-Gibson, Damian J M Tolan, Sophie Whyte, Stewart Bonnington, Alison Scope, Ruth Wong, Barbara Hibbert, John Marsh, and 3 others. 2020. British Society of Gastroenterology/Association of Coloproctology of Great Britain and Ireland/Public Health England post-polypectomy and post-colorectal cancer resection surveillance guidelines. *Gut*, 69(2):201.

Manil Shrestha and Edward Kim. 2026. Conformal Prediction for Risk-Controlled Medical Entity Extraction Across Clinical Domains.

Gyuseon Song, Su Jin Chung, Ji Yeon Seo, Sun Young Yang, Eun Hyo Jin, Goh Eun Chung, Sung Ryul Shim, Soonok Sa, Moongi Simon Hong, Kang Hyun Kim, Eunchan Jang, Chae Won Lee, Jung Ho Bae, and Hyun Wook Han. 2022. Natural Language Processing for Information Extraction of Gastric Diseases and Its Application in Large-Scale Clinical Research. *Journal of Clinical Medicine*, 11(11):2967.

Irena Spasic and Kate Button. 2020. Patient Triage by Topic Modeling of Referral Letters: Feasibility Study. *JMIR Medical Informatics*, 8(11):e21252.

Jonathon Stewart, Juan Lu, Adrian Goudie, Glenn Arendts, Shiv Akarsh Meka, Sam Freeman, Katie Walker, Peter Sprivulis, Frank Sanfilippo, Mohammed Bennamoun, and Girish Dwivedi. 2023. Applications of natural language processing at emergency department triage: A narrative review. *PLOS ONE*, 18(12):e0279953.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):17.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. 2022. MAPIE: an open-source library for distribution-free uncertainty quantification.

Glenn Shafer Vladimir Vovk, Alexander Gammerman. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag, New York.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know Your Limits: A Survey of Abstention in Large Language Models.

Jonathan R. White and Matthew Banks. 2022. Identifying the pre-malignant stomach: from guidelines to practice. *Translational Gastroenterology and Hepatology*, 7:8–8.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings.

Agathe Zecevic, Laurence Jackson, Xinyue Zhang, Polychronis Pavlidis, Jason Dunn, Nigel Trudgill, Shahd Ahmed, Pierfrancesco Visaggi, Zanil YoonusNizar, Angus Roberts, and Sebastian S. Zeki. 2024. Automated decision making in Barrett's oesophagus: development and deployment of a natural language processing tool. *npj Digital Medicine*, 7(1):312.

Xin Zhang, Qiyu Wei, Yingjie Zhu, Fanyi Wu, and Sophia Ananiadou. 2025a. THCM-CAL: Temporal-Hierarchical Causal Modelling with Conformal Calibration for Clinical Risk Prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 916–928, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xinyue Zhang, Agathe Zecevic, Sebastian Zeki, and Angus Roberts. 2025b. Improving Barrett's Oesophagus Surveillance Scheduling with Large Language Models: A Structured Extraction Approach. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 176–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Label Schema and Annotation Rules

Table 5 summarises the six routing categories used for manual annotation of the multi-label task.

Annotation principles. Non-gastrointestinal background conditions were ignored unless they directly changed the interpretation of the procedure. Minor incidental findings that do not typically define a surveillance pathway on their own such as hiatus hernia or inlet patch were not sufficient to trigger *Other* in isolation.

Label	Description
Normal	No abnormal clinical finding identified, procedure is normal
Barrett's	Active Barrett's oesophagus finding
CP	Colorectal polyp finding of any relevant subtype
GA	Gastric atrophy and/or gastric intestinal metaplasia
Other	Active non-target pathology outside the predefined surveillance pathways (for instance: stricture, ulcer, tumour, coeliac-related abnormality)
Incomplete	Incomplete, failed, or abandoned procedure requiring repeat assessment or rebooking

Table 5: Routing label schema used for manual annotation.

Operational rules for recurrent ambiguous cases. The following predefined rules were applied throughout annotation:

- Historical-only mentions were ignored for routing. For example, "history of Barrett's" without evidence of active Barrett's in the current report was not annotated as *Barr*.
- *Normal* was reserved for reports with no pathway-relevant clinical finding and therefore did not co-occur with other labels in the annotations.
- Gastric atrophy and gastric intestinal metaplasia were annotated as *GA*. Co-occurring gastritis alone did not trigger an additional *Other* label.
- Atrophy described outside the stomach was not annotated as *GA*.
- *Incomplete* was assigned for failed, abandoned or incomplete procedures and could co-occur with other active pathway labels when relevant findings were also present.
- *Other* captured active non-target abnormalities outside the predefined surveillance pathways.

B Significance tests

Table 6: Pairwise paired bootstrap significance tests ($n = 1{,}000$ resamples, two-tailed). $\Delta = \text{score}(A) - \text{score}(B)$. $^{*}p < 0.05$, $^{**}p < 0.01$

Model A	Model B	Δ Subset Acc $\uparrow$	Δ Macro F1 $\uparrow$	Δ AURC $\downarrow$
Keyword	TF-IDF	-0.108^{**}	-0.077^{**}	$+0.203^{**}$
Keyword	Embedding	-0.167^{**}	-0.071^{**}	$+0.238^{**}$
Keyword	Fine-tuned	-0.273^{**}	-0.156^{**}	$+0.288^{**}$
TF-IDF	Embedding	-0.059^{**}	$+0.006$	$+0.034^{*}$
TF-IDF	Fine-tuned	-0.165^{**}	-0.079^{**}	$+0.085^{**}$
Embedding	Fine-tuned	-0.106^{**}	-0.085^{**}	$+0.050^{**}$

MedCAT v2: a modular, extensible architecture for clinical named entity recognition and linking under real-world privacy and compute constraints

Mart Ratas[1], Thomas Searle[1], Adam Sutton[1], Richard Dobson[1,2,3],
[1]The Department of Biostatistics & Health Informatics; Institute of Psychiatry, Psychology and Neuroscience; King's College London; London, United Kingdom,
[2]Institute of Health Informatics; University College London; London, United Kingdom,
[3]NIHR Mausley Biomedical Research Centre; London, United Kingdom,
Correspondence: mart.ratas@kcl.ac.uk

Abstract

MedCAT is an open-source framework for clinical named entity recognition and linking (NER+L) widely used in research and healthcare settings. We present MedCAT v2, a re-engineered version designed to improve modularity, extensibility, and maintainability while preserving the core functionality and performance of previous releases. The new architecture introduces a registry-based component system and a flexible pipeline that enables easy substitution of components, integration of alternative methods, and future expansion, including support for pre-trained components across the full NER+L and contextualisation workflow. This enables systematic exploration of clinical NER+L design trade-offs by evaluating different components in the pipeline. Evaluation across multiple public datasets shows equivalent or improved performance compared to earlier versions, with reduced integration overhead and improved runtime flexibility. The framework also supports optional extensions such as meta-annotation, relation extraction, providing a unified and reproducible environment for clinical NLP in real-world settings.

1 Introduction

Clinical free text (e.g., clinic letters, radiology reports, pathology notes, and discharge summaries) contains a substantial proportion of clinically relevant information that is not captured in structured electronic health records (EHR), limiting its secondary use for research, quality improvement, and decision support (Capurro et al., 2014; Kharrazi et al., 2018; Seinen et al., 2025). Advances in clinical natural language processing (NLP) now enable large-scale extraction of biomedical concepts from such data (Elvas et al., 2025). However, deploying such methods in routine clinical and research workflows introduces additional practical constraints: deployments often must run on-premises due to governance and privacy requirements (Van Bulck et al., 2024), may lack GPU-accelerated infrastructure (Wu et al., 2022), and require systems that can be maintained and adapted without substantial engineering overhead (Liu et al., 2025).

MedCAT (Medical Concept Annotation Toolkit) is an open-source framework for clinical named entity recognition and linking (NER+L) [1] that integrates large clinical ontologies such as SNOMED-CT and UMLS(Kraljevic et al., 2021), and also supports contextual meta-annotations (Mascio et al., 2020a; Agarwal et al., 2025b). The toolkit has been deployed across multiple institutions in the UK and internationally, enabling scalable extraction of structured clinical data in real-world settings (Mascio et al., 2020b; Noor et al., 2022; Au Yeung et al., 2024; Van Es et al., 2023).

Despite its success, operational use of MedCAT v1 highlighted several architectural limitations that restricted experimentation with alternative approaches. Exploring alternative approaches often required modifying or forking the codebase, increasing the maintenance burden. In addition, preserving backward compatibility while supporting diverse deployment environments introduced further complexity, particularly in resource-constrained clinical settings.

To address these challenges, we present MedCAT v2, a re-engineered modular architecture designed to support flexible configuration while preserving the proven core functionality of earlier releases. MedCAT v2 introduces a registry-based component system that enables interchangeable tokenizers, NER modules, and linkers, allowing alternative implementations to be integrated without modifying the core pipeline. This design supports systematic exploration of different NER+L configurations and facilitates experimentation with alternative modelling approaches.

[1] https://github.com/CogStack/cogstack-nlp/tree/main/medcat-v2

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 191–198
July 3-4, 2026 ©2026 Association for Computational Linguistics

MedCAT v2 provides a sustainable foundation for clinical NLP by reducing integration overhead and exposing practical trade-offs between accuracy, throughput, and resource use. Our primary contributions are as follows:

- **Modular Architecture**: We introduce a registry-based component system for clinical NER+L that enables interchangeable tokenizers, NER modules, and linkers without modifying the core pipeline.
- **Performance-Throughput Framework**: We provide a redesigned infrastructure that allows for the systematic exploration of deployment trade-offs, supporting both lightweight CPU-only and high-throughput GPU-enabled environments.
- **Empirical Evaluation**: We demonstrate across multiple public datasets that this modular approach maintains or improves upon the performance of previous versions while offering significantly greater flexibility for real-world clinical deployment.

2 MedCAT v2 Architecture

2.1 Limitations of the MedCAT v1 architecture

Fig.1 demonstrates the monolithic architecture built around the spaCy framework. The components of the pipeline operated directly on *spaCy* objects such as documents, tokens, and spans, enabling a straightforward implementation and tight integration with the *spaCy* ecosystem.

However, this design introduced strong coupling between MedCAT's internal components and the underlying *spaCy* data structures. Changes to upstream processing steps, such as tokenization, required corresponding modifications throughout the rest of the pipeline to ensure compatibility with the expected document and span representations.

In addition, MedCAT v1 did not provide a mechanism for defining or registering alternative implementations of tokenizers or core components (e.g named entity recognition (NER) or entity linking (EL)). Exploring alternative approaches therefore required modifying or forking the codebase and integrating new implementations directly into the pipeline logic, increasing maintenance overhead and making experimentation with alternative methods difficult.

2.2 Modular and extensible pipeline architecture

To address these limitations, MedCAT v2 adopts a modular pipeline architecture designed around interchangeable components. Core stages of the pipeline (tokenization, named entity recognition (NER), and entity linking (EL)) are independently defined and registered, allowing alternative implementations to be integrated without modifying the rest of the system.

A lightweight *registry system* manages the available components and their implementations. This enables components to be swapped, extended, or replaced while maintaining a consistent pipeline interface. In addition to the core components, MedCAT v2 supports optional add-on modules that operate on the annotated document, such as MetaCAT for contextual meta-annotations and RelCAT for relation extraction (Agarwal et al., 2025a).

The registry architecture also enables support for external extensions: independent software packages that provide additional component implementations. These extensions can register their components with the MedCAT pipeline at runtime, allowing new functionality to be distributed separately from the core framework. For example, alternative linking approaches such as embedding-based linkers can be provided as external extensions while remaining fully compatible with the MedCAT pipeline.

To support reproducibility and deployment transparency, the framework tracks the provenance of each component used in a pipeline, including whether it originates from the core framework or from an external extension. This information is recorded in the model metadata, allowing pipelines to be fully described and reconstructed across different environments.

Fig. 2 describes the high level architecture of MedCAT 2.0. Components not explicitly discussed in this paper retain the same underlying implementation and behaviour as in MedCAT v1, and readers are referred to the original MedCAT publications for additional details on these components.

2.3 Example component implementations

The registry-based architecture allows different implementations of each pipeline stage to be configured depending on the requirements of a deployment. Components can be provided either by the core MedCAT library or by external extension pack-

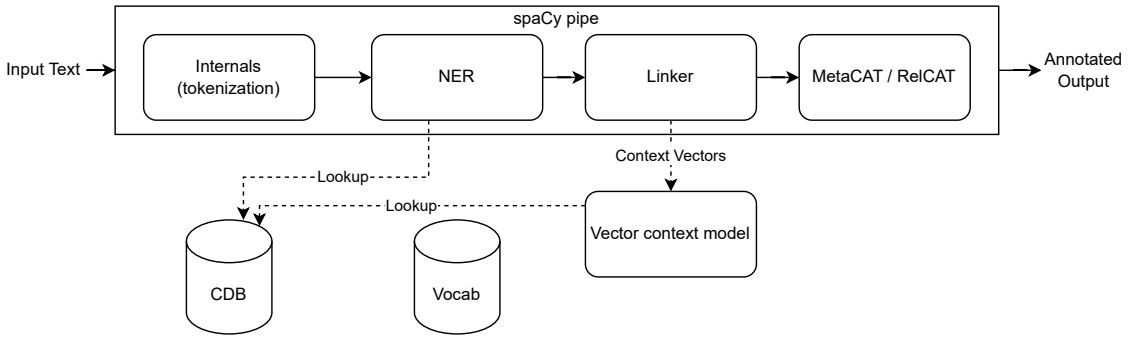


Figure 1: High-level architecture of MedCAT v1 showing the monolithic pipeline design built around *spaCy*.

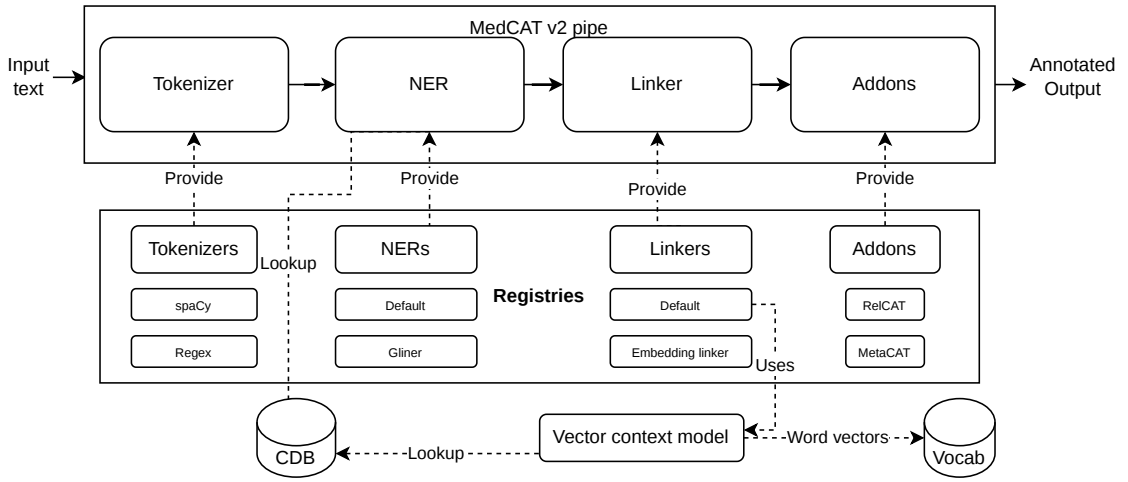


Figure 2: Modular registry-based architecture of MedCAT v2, illustrating interchangeable core components, optional add-on modules, and example registered implementations.

ages, enabling new functionality to be integrated without modifying the core system. For example, tokenization can be performed using either the default *spaCy*-based tokenizer or a lightweight regular-expression tokenizer designed for minimal environments. Similarly, alternative implementations of the NER and entity linking components can be used to balance trade-offs between computational cost, throughput, and predictive performance. To illustrate this flexibility, we highlight several component implementations available with MedCAT v2.

We demonstrate the improved flexibility through the following new components:

- **Regex-based tokenizer.** A lightweight tokenizer based on regular expressions designed for high-throughput or resource-constrained environments. While this approach reduces dependency overhead and improves processing speed, it may sacrifice some linguistic accuracy compared to more sophisticated tokenization methods
- **Faster linker.** A heuristic linker that resolves entities when a surface form maps unambiguously to a single concept, or when only one candidate concept treats the surface form as its primary name. By bypassing full disambiguation, this component substantially improves inference speed at the cost of reduced recall in ambiguous contexts.
- **Embedding-based linker.** A disambiguation component that leverages transformer based architectures (e.g. *sentence-transformers/all-MiniLM-L6-v2*) to improve concept selection for short or ambiguous mentions. In the configuration evaluated in this work, the linker uses static embeddings and provides modest improvements depending on the dataset and configuration. However, the architecture also supports training task-specific embedding representations for entity disambiguation, which

can substantially improve performance. This implementation is distributed as an independent MedCAT extension package on PyPI, demonstrating how external components can provide alternative implementations for core pipeline stages.

- *Gliner based NER*. A named entity recognition component based on GLiNER (Zaratiana et al., 2024), which can be used as an alternative to the default MedCAT vocabulary-driven NER pipeline. This component demonstrates how transformer-based external NER systems can be integrated into MedCAT v2 through the modular pipeline architecture. While the default MedCAT approach relies on vocabulary matching combined with normalization and spell-correction mechanisms, the GLiNER-based implementation enables experimentation with fully neural NER approaches within the same downstream linking pipeline. This is an NER step based on gliner (Zaratiana et al., 2024).

2.4 Backward compatibility and migration

To support a smooth transition for existing users, MedCAT v2 includes a legacy compatibility layer that allows models trained with MedCAT v1 to be loaded and used within the new framework with no user intervention. This ensures that previously trained models and annotation efforts can be preserved while benefiting from the improved architecture and modular pipeline introduced in v2. In addition, we provide a migration pathway and accompanying documentation to assist users in transitioning existing workflows to the new system (Ratas, 2026).

3 Results

3.1 Overall NER+L performance

MedCAT v2 maintains equivalent performance to MedCAT v1 when using the same trained model and default vector context model for entity linking. Across all evaluated datasets, precision, recall, and F1-scores remain within the expected variance range using a private internal model[2]. To enable direct comparison, equivalent pipeline configurations and model parameters were used across both versions wherever possible. These results confirm that the redesigned modular architecture does not adversely affect core NER+L functionality. A detailed comparison between the two versions is shown in Table 1.

To evaluate performance parity between versions, we evaluated both systems across several biomedical NER+L datasets including MDACE (Cheng et al., 2023), COMETA (Basaldella et al., 2020), Distemist (Miranda-Escalada et al., 2022), the SNOMED Linking Challenge (Davidson et al., 2025), and MedMentions (Mohan and Li, 2019). All datasets were preprocessed into a common evaluation format in order to obtain comparable metrics.

3.2 Runtime performance

We compared MedCAT v1 and v2 across several common operational scenarios, including model loading, inference, and unsupervised training. Across all experiments, runtime performance remains broadly comparable between versions. Inference latency is effectively unchanged when equivalent configurations are used, indicating that the architectural refactor does not introduce additional overhead during normal pipeline execution. Model loading times show modest variation, with v2 slightly faster in some cases, while unsupervised training times remain within the expected range of implementation and hardware variability. A summary of timing benchmarks is presented in Table 2. 20 MIMIC-IV documents were used within the inference unsupervised training runs depicted in the table. All experiments were run on a single machine with an Apple M1 CPU and 32 GB RAM.

To ensure comparability, all timing experiments were performed under identical conditions and with the same base model configuration. The results therefore reflect differences introduced by architectural changes alone, rather than variations in model structure or parametrisation.

3.3 Performance–throughput trade-offs across configurations

A key advantage of the MedCAT v2 architecture is that it exposes multiple interchangeable components, allowing users to configure the pipeline to match their desired balance between predictive performance and processing speed. This flexibility is particularly important across the diverse range of MedCAT use cases, from large-scale processing of millions of EHRs across extensive concept vocabularies to targeted extraction tasks involving a small number of concepts.

[2]This model is trained on NHS data and therefore cannot be publicly released.

Table 1: Comparison of MedCAT v1 and MedCAT v2 performance across multiple biomedical NER+L datasets using a private internal model. Differences between versions are shown in the "diff." columns.

Dataset	Precision			Recall			F1		
	v1	v2	diff.	v1	v2	diff.	v1	v2	diff.
MDACE	0.3656	0.3660	0.0004	0.5391	0.5398	0.0007	0.4357	0.4362	0.0005
COMETA	0.9242	0.9245	0.0003	0.4521	0.4521	0.0000	0.6072	0.6072	0.0001
Distemist	0.3187	0.3192	0.0005	0.3949	0.3958	0.0009	0.3527	0.3534	0.0007
Linking Challenge	0.5353	0.5353	0.0000	0.3338	0.3337	-0.0001	0.4112	0.4112	0.0000
Med-Mentions	0.3521	0.3521	0.0000	0.5104	0.5103	0.0001	0.4167	0.4167	0.0000

Table 2: Runtime comparison between MedCAT v1 and v2 across inference, model loading, and unsupervised training scenarios. MC refers to MetaCAT components included in the pipeline.

Scenario	Ver	Model	Time - v1 (s)	Time - v2 (s)
inference	v1	NER (no MC)	9.090	10.827
inference	v1	NER + 3 MCs	22.537	22.311
load	v1	NER (no MC)	12.030	12.061
load	v1	NER + 3 MCs	12.218	17.653
unsup_train	v1	NER (no MC)	12.121	10.147
unsup_train	v1	NER + 3 MCs	22.392	26.776

On COMETA, the lightweight components deliver substantial runtime gains with only modest reductions in F1. For example, the combination of the regex tokenizer and the faster linker can achieve more than an order-of-magnitude speedup relative to the standard configuration, while maintaining comparable precision and only a small reduction in recall. This suggests that for some large-scale or latency-sensitive applications, MedCAT v2 can operate at significantly higher throughput without materially compromising output quality.

In contrast, results on the Linking Challenge dataset highlight that no single configuration is universally optimal. While the faster linker again provides clear speed improvements, the regex tokenizer does not produce a similar benefit. This likely reflects differences in dataset structure: COMETA consists of smaller, tightly filtered projects, whereas the Linking Challenge dataset allows a much larger variety of candidate concepts. As a result, the regex tokenizer produces many more potential mentions that must be evaluated during linking, increasing overall processing time. These results illustrate that tokenization and linking behaviour can interact with dataset characteristics in non-trivial ways.

The embedding-based linker represents the high-resource end of this configuration spectrum. While it is designed to improve disambiguation for short or ambiguous mentions, it introduces substantially higher computational cost and typically requires GPU acceleration for practical use. Across the evaluated datasets it does not consistently outperform lighter linking strategies, sometimes offering only marginal improvements despite significantly longer runtimes. This suggests that contextual embedding approaches are most beneficial in settings where ambiguity is a dominant challenge, but provide less advantage when candidate sets are already well constrained.

Overall, these experiments demonstrate that MedCAT v2 is intentionally flexible rather than prescriptive. Instead of enforcing a single configuration, the system allows practitioners to select pipeline components that match their operational constraints, whether prioritising throughput, resource efficiency, or predictive performance. The variation across datasets also highlights the importance of validating component choices on task-specific data rather than assuming a universally optimal configuration.

4 Discussion

The goal of this work was to address practical deployment and maintenance challenges encountered with MedCAT in real-world clinical and research

Table 3: Performance and runtime for different MedCAT v2 configurations across two datasets. Embedding linker results were obtained using the *abhinand/MedEmbed-small-v0.1* model

Dataset	Configuration	Precision	Recall	F1	Time (s)
COMETA					
Spacy	Vector context	0.9245	0.4521	**0.6072**	68.16
Spacy	Faster linker	**0.9266**	0.4225	0.5804	51.64
Spacy	Embedding linker	0.8871	**0.4455**	0.5932	321.37
Regex	Vector context	0.9130	0.4136	0.5693	30.54
Regex	Faster linker	0.9205	0.4108	0.5681	**6.21**
Regex	Embedding linker	0.8759	0.4394	0.5852	348.79
2023 Linking Challenge					
Spacy	Vector context	0.5353	0.3337	0.4112	75.40
Spacy	Faster linker	**0.5934**	0.2873	0.3871	**48.05**
Spacy	Embedding linker	0.5171	0.3557	**0.4215**	511.19
Regex	Vector context	0.4522	0.3162	0.3722	117.55
Regex	Faster linker	0.5091	0.2862	0.3664	82.61
Regex	Embedding linker	0.4141	**0.3593**	0.3847	248.02

environments, while preserving its established performance. Our results show that MedCAT v2 maintains equivalent predictive behaviour to earlier releases while substantially increasing flexibility through a modular, registry-based architecture.

Operational experience with MedCAT v1 highlighted recurring challenges: tightly coupled components made it difficult to experiment with alternative models, upgrades often required invasive code changes, and deployment across heterogeneous infrastructure introduced avoidable complexity. MedCAT v2 directly targets these issues by decoupling core pipeline elements and enabling components to be composed, replaced, or omitted without modification of the core library, and allowing new functionality to be integrated through external extensions. In practice, this allows development teams to integrate new tokenizers, linkers, or contextual classifiers independently, while clinical informatics groups can deploy lightweight CPU-only configurations tailored to local constraints. The extension system enabled external packages to provide components for MedCAT without cluttering up the core library.

The redesigned framework also exposes explicit trade-offs between accuracy, throughput, and resource consumption. Our experiments demonstrate that simplified components can yield substantial gains in processing speed, whereas embedding-based linking can improve disambiguation in some contexts but at substantially higher computational cost. These findings reinforce that no single configuration is optimal across deployments, and that teams should evaluate component choices against their own corpora, performance requirements, and infrastructure availability rather than relying solely on aggregate benchmark metrics.

Clinical NLP systems span a wide spectrum, from lightweight rule-based pipelines to fully neural end-to-end architectures. MedCAT v2 is designed to occupy a practical middle ground, supporting constrained environments while remaining extensible to more resource-intensive components where applicable. Rather than replacing specialised systems, the framework aims to lower the barrier to deployment and experimentation in settings where operational considerations often dominate model selection.

Several limitations remain. Embedding-based linking is computationally demanding and may be impractical without GPU access. Performance continues to depend on ontology coverage and data availability, with openly available models lagging institution-specific deployments. Increased configurability also introduces a larger design surface for users. To mitigate this, MedCAT v2 is accompanied by migration tooling, documentation, tutorials, and reference configurations, with ongoing refinement guided by community feedback.

Overall, MedCAT v2 provides a more sustainable engineering foundation for clinical concept extraction. By reducing operational friction while retaining established performance, the framework supports broader adoption of clinical NLP across

research and applied healthcare environments.

5 Conclusions

MedCAT v2 delivers a more operationally practical foundation for clinical named entity recognition and linking, preserving the established performance of earlier releases while substantially improving deployment flexibility and maintainability. By decoupling core components and simplifying integration, the framework reduces engineering overhead and enables teams to adapt pipelines to their local infrastructure and use-case requirements.

The redesigned architecture supports a wide range of deployment profiles, from lightweight CPU-only configurations within hospital infrastructure to more advanced pipelines that incorporate embedding-based components. The modular design also enables external extensions to provide alternative implementations of core components without requiring changes to the framework itself. Crucially, backward compatibility ensures continuity for existing users: previously trained models and data formats can be reused without modification, protecting prior investments while enabling incremental adoption of new functionality.

As an open-source platform, MedCAT v2 (Ratas et al., 2026) provides a sustainable basis for community-driven development and applied clinical NLP. The modular design positions the toolkit to integrate emerging methods in linking, contextualisation, and privacy-preserving workflows without disruptive architectural changes. By lowering barriers to experimentation, deployment, and long-term maintenance, MedCAT v2 supports more robust translation of NLP research into real-world healthcare environments and enables scalable extraction of structured clinical information from unstructured text.

References

Shubham Agarwal, Vlad Dinu, Thomas Searle, Mart Ratas, Anthony Shek, Dan F. Stein, James Teo, and Richard Dobson. 2025a. RelCAT: Advancing extraction of clinical inter-entity relationships from unstructured electronic health records. *Preprint*, arXiv:2501.16077.

Shubham Agarwal, Thomas Searle, Mart Ratas, Anthony Shek, James Teo, and Richard Dobson. 2025b. A framework for flexible extraction of clinical event contextual properties from electronic health records. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 946–959, Vienna, Austria. Association for Computational Linguistics.

Joshua Au Yeung, Anthony Shek, Thomas Searle, Zeljko Kraljevic, Vlad Dinu, Mart Ratas, Mohammad Al-Agil, Aleksandra Foy, Barbara Rafferty, Vitaliy Oliynyk, and 1 others. 2024. Natural language processing data services for healthcare providers. *BMC medical informatics and decision making*, 24(1):356.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Daniel Capurro, Meliha Yetisgen, Erik Eaton, Robert Black, and Peter Tarczy-Hornoch. 2014. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: A multi-site assessment. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, Gleb Sokolov, and Chen Yanover. 2025. Snomed ct entity linking challenge. *Journal of the American Medical Informatics Association*, 32(9):1397–1406.

Luis B. Elvas, Ana Almeida, and João C. Ferreira. 2025. Natural language processing in medical text processing: A scoping literature review. *International Journal of Medical Informatics*, 204:106049.

Hadi Kharrazi, Laura J Anzaldi, Leilani Hernandez, Ashwini Davison, Cynthia M Boyd, Bruce Leff, Joe Kimura, and Jonathan P Weiner. 2018. The value of unstructured electronic health record data in geriatric syndrome case identification. *J. Am. Geriatr. Soc.*, 66(8):1499–1507.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Leibo Liu, Victoria Blake, Matthew Barman, Blanca Gallego, Timothy Churches, Georgina Kennedy, Sze-Yuan Ooi, Geoffrey P Delaney, and Louisa Jorm. 2025. Using natural language processing to extract information from clinical text in electronic medical records for populating clinical registries: a systematic review. *Journal of the American Medical Informatics Association*, page ocaf176.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020a. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 86–94, Online. Association for Computational Linguistics.

Aurelie Mascio, Željko Kraljević, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020b. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th sigbiomed workshop on biomedical language processing*, pages 86–94.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Kawsar Noor, Lukasz Roguski, Xi Bai, Alex Handy, Roman Klapaukh, Amos Folarin, Luis Romao, Joshua Matteson, Nathan Lea, Leilei Zhu, and 1 others. 2022. Deployment of a free-text analytics platform at a uk national health service research hospital: Cogstack at university college london hospitals. *JMIR medical informatics*, 10(8):e38122.

Mart Ratas. 2026. MedCAT v2 migration guide. `https://github.com/CogStack/cogstack-nlp/blob/main/medcat-v2/docs/migration_guide_v2.md`. Accessed: 2026-04-24.

Mart Ratas, Zeljko Kraljevic, Anthony Shek, Thomas Searle, and Xi Bai. 2026. Medcat v2.

Tom M Seinen, Jan A Kors, Erik M van Mulligen, and Peter R Rijnbeek. 2025. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *J Med Internet Res*, 27:e66910.

Liesbet Van Bulck, Meghan Reading Turchioe, Maxim Topaz, and Jiyoun Song. 2024. Exploring the full potential of the electronic health record: the application of natural language processing for clinical practice. *European Journal of Cardiovascular Nursing*, 24(2):332–337.

Bram Van Es, Leon C Reteig, Sander C Tan, Marijn Schraagen, Myrthe M Hemker, Sebastiaan RS Arends, Miguel AR Rios, and Saskia Haitjema. 2023. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, 24(1):10.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael TC Poon, Natalie Fitzpatrick, Adam P Levine, and 1 others. 2022. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *NPJ digital medicine*, 5(1):186.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Effects of Adaptive Pretraining in Specialized Domains for Named Entity Recognition

Jack Lynam
Christopher Newport University
1 Avenue of the Arts
Newport News, Virginia, USA 23606
jack.lynam@cnu.edu

Sam Henry
Randolph-Macon College
114 College Ave
Ashland, Virginia, USA 23005
samuelhenry@rmc.edu

Abstract

Due to unique concepts, syntactic structure, and vocabulary of specialized domains, it is common to train specialized Language models (LMs) for their target domain. For example, BioClinicalBERT is a specialized LM designed for clinical applications. These specialized LMs are typically created starting with a foundation model (such as BERT-base) which has been pretrained for the general English domain, and then adapted to the target domain via additional pretraining. Alternatively, LMs may be pretrained from scratch on data from the target domain. Both techniques are extremely computationally expensive and as such, these specialized LMs are often publicly released for other researchers. For some domains, such as the biomedical domain there are many, similar models available, and as a developer, this raises the question, which pretrained LM should I choose? Alternatively, in novel domains for which no specialized LMs exist, it raises different questions: Is it worth the cost to pretrain a LM from scratch? Should I adapt a general English model instead? Should I just use a general English model without adaptive pretraining? This is a particularly salient question when considering a limited budget. i.e. Should I pay for compute time or for annotators to create a larger dataset. In this paper we compare results of nine LMs across nine datasets spanning the clinical, scientific, and biomedical-related social media domains. From these comparisons we make several conclusions that can simplify the hyperparameter-tuning process and inform researchers and developers in novel domains. Broadly, these are that the effects of adaptive fine-tuning are small. If an adapted model exists in your domain, choose the one most closely related to your task. If no model exists, using a foundation model is likely sufficient.

1 Introduction

Named Entity Recognition (NER) is the task of identifying salient terms in text, and is a core NLP task performed within these larger systems, and as such, improving NER is a vital step to improving their effectiveness. Transformer (Vaswani et al., 2017)-based language models (LM) have become the state-of-the art for many NLP tasks, and despite recent advancements in Large Language Models (LLMs), smaller Language Models (LMs) such as BERT (Devlin et al., 2018) remain the state-of-the-art for low-level tasks such as NER. Although smaller than the larger LLMs, these models are still extremely computationally expensive to train from scratch and as such are typically created for maximum utility within the General English domain. However, specialized domains such as the biomedical and clinical domains have unique concepts, language, and grammatical structure (Nadkarni et al., 2011) meaning these General English models may not perform optimally without domain-specific adaptation.

Evidence suggests that BERT models perform better when pretrained on data closely related to their target task (Lee et al., 2020; Alsentzer et al., 2019; Gu et al., 2021; Peng et al., 2019), and previous work has compared several LMs across multiple datasets. For example, Peng et al propose the Biomedical Language Understanding Evaluation (BLUE) benchmark and compare the effects of pretraining on PubMed and PubMed + MIMIC using Large and Small BERT models. Gu et al (Gu et al., 2021) propose the Biomedical Language Understanding & Reasoning Benchmark (BLURB) benchmark and compare six biomedical language models and Base BERT over various tasks including five NER datasets. These works serve as important biomedical benchmarks to compare biomedical LMs. We distinguish ourselves from these works in focusing on understanding the effects of adaptive pretraining to answer practical questions such as, what LM should I choose or is it worth it to adaptively pretrain an LM for a novel domain? To answer this, we collected and compared results of

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 199–208
July 3-4, 2026 ©2026 Association for Computational Linguistics

nine different LMs over nine popular NER datasets in the clinical, scientific, and biomedical-related social media domains.

We find that while domain-specific pretraining does increase performance, the increase is less dramatic than expected, not across-the-board, and often not statistically significant. This is even true for LMs trained from scratch for the target domain. Given the extremely high computational costs, these results have important implications for researchers working in novel domains without pre-existing domain-specific LMs, and can help guide the choice of a LM for researchers and developers in domains with pre-existing domain-specific LMs.

2 Background

2.1 Training Language Models

LMs consist of an Encoder-Decoder architecture. In this architecture, the encoder learns effective numeric representations of the text, and the decoder transforms those numeric representations into an output useful for the target task. LMs are trained in a sequence of steps: (1) *Pretraining* - train the encoder to create numeric representations of text. (2) *Optional adaptive pretraining* - adjust those numeric representations for the target domain. (3) *Fine-tuning* - train a decoder (and usually also adjust the weights of the encoder) to perform well on a specific task and dataset.

Pretraining is a self-supervised step in which the LM is trained on massive corpora containing billions of words. The precise pretraining routine varies depending on the specific LM (Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020; He et al., 2020), however most follow a routine similar to BERT (Devlin et al., 2018). BERT is jointly trained using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is performed by masking words in a sentence, and using the LM to predict the masked word. NSP is performed by providing two text samples to BERT, and training it to predict whether the second sentence follows the first sentence.

Pretraining is by far the most computationally expensive step of LM development, and creating a LM from scratch often carries prohibitive costs. As such, pretrained LMs are typically created for maximum utility and therefore trained for the general English domain. These pretrained LMs are often released as foundation models which are used by researchers directly or adapted to other datasets via adaptive pretraining.

In the *optional adaptive pretraining* step, the LM is further pretrained (typically using the same pretraining routine) on domain-specific corpora. For example, BERT-Base which is pretrained for the general English domain may adapted to the biomedical domain by performing additional pretraining using PubMed abstracts. While less computationally expensive than initial pretraining, this is still extremely computationally expensive (often prohibitively expensive) and the adapted LMs are commonly released for other developers in their domain.

Fine-tuning is the last step in training a LM. In this step, a task-specific decoder is attached to the pretrained weights of the encoder portion of the LM and the full model is trained to maximize performance on the provided dataset. This is often the only step performed by developers, who will select a pretrained LM using online repositories such as HuggingFace[1], then train it on a specific dataset. Common wisdom suggests you select a LM most closely related to your target domain.

2.2 Biomedical Data Sources

Publicly available databases have been created to enable research in biomedicine. Two of the most popular text databases are PubMed, which contains biomedical literature from online journals and books, and MIMIC-III, which contains standardized, de-identified clinical records.

The PubMed (Wheeler et al., 2007) and PubMedCentral (Roberts, 2001) resources provided by the National Center for Biotechnology Information (NCBI) are repositories of scientific articles related to the biomedical sciences. PubMed contains over 40 million article abstracts while PubMedCentral contains nearly 12 million full-text articles.

The Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016) is a vast, freely-available database containing about 2 million de-identified clinical records. MIMIC-III data is sourced from critical care reports at Beth Israel Deaconess Medical Center, a hospital in Boston, Massachusetts. These clinical records report on a variety of topics that include laboratory testing, billing and administrative work, patient demographics, discharge summaries, radiology reports, chart write-ups, and pharmacotherapy.

[1] https://huggingface.co/

Domain	Model	General English		Scientific		Clinical		Social Media	
		Wiki	Books Corpus	PubMed full-text	PubMed abstracts	MIMIC all-types	MIMIC discharge	Twitter	Reddit
General English	BERT	x	x						
Scientific	Bi-B	x	x	x	x				
	Blue-B	x	x		x				
	PMed-B				x				
Clinical	BiClin-B	x	x	x	x	x			
	BiDis-B	x	x	x	x		x		
	BlueB+M	x	x		x	x			
Social Media	BT	x	x					x	
	BiRed-B	x	x	x	x				x

Table 1: Language models and their pre-training data sources.

2.3 Pretrained Language Models for the Biomedical Domain

The biomedical domain is a broad, well-studied domain and as such has many pretrained LMs available for download. We divide the biomedical domain into three main sub-domains: (1) *clinical* which relates to the medical aspects of biomedicine, (2) *scientific* which relates to biology, chemistry, genetic, and other scientific aspects of biomedicine, and (3) *social media* which relates to consumer and patient views and testimony related to biomedicine. For these experiments, we selected nine different LMs which cover these sub-domains. All selected LMs except PubMedBERT are domain-specific adaptations of the general English BERT-Base model, and all but BERT Tweet are pretrained in the same way as BERT (e.g. MLM and NSP). Uniquely, PubMedBERT is trained from scratch for the scientific biomedical domain. We describe each evaluated LM below. To make the tables more compact, we assigned abbreviations for each of these LMs, which are listed in bold. Table 1 summarizes their difference in training data sources.

BERT: BERT-Base is the original, general English model proposed by Devlin et al (Devlin et al., 2018). It was pre-trained for the MLM and NSP tasks using the Books Corpus (Zhu et al., 2015) and English Wikipedia.

Bi-B: BioBERT (Lee et al., 2020) is initialized from BERT-Base and adaptively pre-trained on PubMed abstracts and PubMed full-text articles from PubMed Central. In the biomedical domain, Bi-B has significantly outperformed BERT on several tasks including NER and relationship extraction (Lee et al., 2020; Gu et al., 2021), and was therefore used as the initial weights for the next two systems.

BiClin-B: BioClinicalBERT(Alsentzer et al., 2019) is initialized from Bi-B and further pre-trained on all clinical notes in the MIMIC-III database. BiClin-B has been shown to be slightly more effective for clinical tasks than Bi-B (Alsentzer et al., 2019).

BiDis-B: BioDischargeSummaryBERT (Alsentzer et al., 2019) is initialized from BioBERT and further pre-trained on clinical notes which are discharge summaries from MIMIC-III. BiDis-B shows improvement over Bi-B for tasks with corpora that deal specifically with discharge summaries (Alsentzer et al., 2019).

Blue-B: BlueBERT (PubMed) (Peng et al., 2019) is initialized from BERT-Base and further pre-trained on PubMed abstracts. Blue-B pre-training data (PubMed) is similar to that of Bi-B, but excludes full-text articles. Blue-B has been shown to perform similarly to Bi-B (Peng et al., 2019).

BlueB+M: BlueBERT (PubMed+MIMIC) (Peng et al., 2019) is initialized from BERT-Base and further pre-trained on PubMed abstracts and additionally clinical notes from MIMIC-III. Therefore, this LM's pre-training data is similar to that of BiClin-B, but excludes full-text PubMed articles. BlueB+M has been shown to outperform Bi-B on clinical tasks (Peng et al., 2019).

PMed-B: PubMedBERT (Gu et al., 2021) is pre-trained from scratch on PubMed abstracts. Unlike the other LMs it is not adapted from BERT Base. Theoretically, this should be hugely beneficial because the vocabulary of a LM is fixed and is determined from the initial pretraining corpus. Whereas LMs adapted from general English may need to represent specialized terms using subword tokens, PMed-B can represent specialized terms

natively because they are present in its specialized vocabulary. PMed-B was shown to perform comparably or better than LMs like Bi-B and Blue-B for biomedical tasks (Gu et al., 2021).

BT: BERTweet (Nguyen et al., 2020) is initialized from BERT Base and further pre-trained on Twitter data. BT is not adapted to the biomedical domain, but is instead adapted to the unique, short-form, highly abbreviated nature of Tweets. We include BT to try to determine what is more important, social media structure or the biomedical context. Instead of following the original BERT pre-training procedure, BT's adaptive pre-training is modeled after the RoBERTa procedure (Liu et al., 2019). This pre-training procedure is distinct in two ways: masked tokens are changed each epoch during MLM and no NSP is performed. BT was shown to outperform similar LMs for NER in the general social media domain (Nguyen et al., 2020).

BiRed-B: BioRedditBERT (Basaldella et al., 2020) is initialized from Bi-B and further pre-trained on text from health-related Reddit threads. BiRed-B has shown good performance for biomedical text mining in the social media domain (Basaldella et al., 2020).

3 Datasets

We selected nine datasets which, like the selected LMs span the clinical, scientific, and social media subdomains of biomedicine. The datasets are summarized in Table 2 and brief descriptions are provided below.

i2b2 2010: The i2b2 2010 Concepts, Assertions, and Relations challenge (Uzuner et al., 2011) aimed to create suitable NLP systems for the clinical domain. Its associated dataset consists of 426 de-identified clinical notes from Partners Healthcare and Beth Israel Deaconess Medical Center. For the NER challenge, the clinical notes were annotated for medical problems, treatments, and tests. The challenge organizers provide a split of 170 training notes and 256 test notes.

n2c2 2018: The n2c2 2018 Track 2 challenge (Henry et al., 2020) focused on developing systems to recognize adverse drug events (ADEs). The dataset consists of 505 de-identified clinical notes gathered from the MIMIC-III clinical care database. For the NER challenge, the data was annotated for entities related to medications and ADEs which include: drug, strength, form, dosage, frequency, route, duration, reason, and ADE. The challenge organizers provide a split of 303 training notes and 202 testing notes.

BC5CDR: The BioCreative V Track 3 Chemical-Disease Relation (CDR) challenge (Wei et al., 2016) was created to develop systems that identify chemical-disease interactions. The CDR corpus is composed of 1500 PubMed abstracts annotated for diseases and chemicals. The task organizers provide a split of 500 training, 500 validation, and 500 test abstracts.

BC7DCPI: The BioCreative VII Track 1 drug-protein and chemical-protein interactions challenge (Miranda et al., 2021) focused on identifying biochemical entities for drug discovery. This dataset is a collection of 5000 PubMed abstracts annotated for genes and chemicals. The organizers provide a split of 3500 training abstracts, 750 validation abstracts, and 750 test abstracts. Background and large-scale background collections of automatically-annotated abstracts were also released for further analysis.

NCBI: The NCBI disease corpus (Dogan and Lu, 2012) is a collection of PubMed abstracts designed to test the efficacy of disease recognition systems. In total, 793 PubMed abstracts were annotated for disease mentions. While the corpus maintainers do not provide a training-test split for the data, one provided by (Crichton et al., 2017) has seen use in biomedical benchmarks (Gu et al., 2021).

NLMChem: The NLM-Chem corpus (Islamaj et al., 2021) was created for the BioCreative VII Track 2 task on identifying chemical information from full-text scientific publications. It contains 150 full-text PubMed articles annotated for chemical mentions. The corpus maintainers provide a split of 80 training documents, 20 validation documents, and 50 test documents.

BC7Med: The BioCreative VII Track 3 challenge (Weissenbacher et al., 2021) focused on identifying medication and dietary supplement mentions from social media data. The dataset contains tweets from 212 different users' timelines annotated for drug name mentions. This dataset is large, yet heavily imbalanced with only 0.2% percent of the tweets containing medication mentions. The challenge organizers provide a split of 89,004 tweets for training, 38,149 tweets for validation, and 54,482 tweets for the test set.

COMETA: The Corpus of Online Medical EntiTies (COMETA)(Basaldella et al., 2020) contains biomedical entities from Reddit. It was originally intended to evaluate entity linking systems and con-

Dataset	Domain	Entity Types	Data Source
BC5CDR	Scientific	Disease, Chemical	PubMed abstracts
BC7DCPI	Scientific	Gene, Chemical	PubMed abstracts
NCBI	Scientific	Disease	PubMed abstracts
NLMChem	Scientific	Chemical	Full-text PubMed articles
i2b2 2010	Clinical	Problem, Treatment, Test	De-identified clinical notes
n2c2 2018	Clinical	Drug, Strength, Form, Dosage, Frequency, Route, Duration, Reason, ADE	De-identified clinical notes
BC7Med	Social Media	Medication	Twitter data
COMETA	Social Media	Miscellaneous Biomedical Entities	Reddit threads
ADEMiner	Social Media	ADE	Twitter data

Table 2: Description of datasets and their sources used in evaluation.

tains 20,015 health-themed posts from Reddit annotated for biomedical entities found within the SNOMED CT ontology (Donnelly et al., 2006). These entities represent a wide variety of biomedical concepts including symptoms, diseases, chemicals, genes, procedures, and more. To re-purpose this dataset for NER, we identified the text spans of the annotated entities and used them as the gold standard. Unique to this dataset (due to its repurposing), all entities within this dataset received a label of "Biomedical Entity" or None.

ADEMiner: The DeepADEMiner corpus (Magge et al., 2021) is a combination of three different datasets (SMM4H 2017 (Sarker et al., 2018), SMM4H 2019 (Weissenbacher et al., 2019), and SMM4H 2020 (Gonzalez et al., 2020)) created for identifying and normalizing ADEs from Twitter data. There are a total of 33,246 tweets that are annotated for ADE mentions. No official training-testing split is provided.

4 Methodology

4.1 Data Processing

LM training requires a maximum of 512 token length samples, so each dataset was processed (chunked) into samples. For some datasets, a sample was clearly defined. This is the case for i2b2 2010, BC7Med, COMETA, and ADEMiner, where annotations are contained on one line and therefore a sample is identified as a single line of text. For other datasets, there is no clear distinction between samples because the text is annotated on a character-level basis and no pre-defined sample deliminator is provided. For these datasets, which include n2c2 2018, BC5CDR, BC7DCPI, NCBI, and NLMChem, we processed these datasets into samples using the ScispaCy sentence splitter (Neumann et al., 2019) with the `en_core_sci_sm` model [2].

[2] https://allenai.github.io/scispacy/

All datasets were tokenized using each LM's specific tokenizer (as specified by HuggingFace's `Transformers.Autotokenizer` library).

4.2 Language Model Architecture and Training

NER was posed as a multi-class token classification problem and performed by predicting a set of labels for each token token in that sample. The set of labels corresponds to the entity types of that dataset plus a *None* class. The `argmax` is taken over the set of labels to convert to a one-hot label encoding for each token. For example, in the BioCreative 5 CDR dataset, a vector of `<Disease, Chemical, None>` probabilities is predicted for each token, and converted to a vector containing all zeros except for a single one indicating the class.

We first performed a preliminary hyperparameter search to determine the learning rate, batch size, embedding dropout, and decoder architecture. Manual analysis of these results and those reported in the literature (Devlin et al., 2018; Alsentzer et al., 2019; Michalopoulos et al., 2020) support that fine-tuning with a learning rate of 1e-5, batch size of 16, and no embedding dropout produces LMs that perform well for biomedical NER. Due to the prohibitive costs of performing separate hyperparameter searches for all LMs and all datasets, we reused these hyperparameters for each LM across every dataset. For similar reasons, we also fixed the neural network architecture across all datasets. All BERT models have a distinct task-specific decoder, meaning they predict only classes found within a particular dataset. The decoder is a single densely-connected layer with softmax activation.

4.3 Results Generation

Reported results are the micro-averaged F1 scores averaged over 5-folds of cross-validation. For datasets that provided training, test, and/or vali-

dation splits, we combined all data into a single set which we used for cross-validation. Each fold consists of 70% for training, 10% for early stopping and 20% for validation (testing). The data-split was randomized but consistent across all compared LMs. To maximize the number of training samples per fold, we used the early stopping split to determine the optimal number of training epochs, then we combined the training and early-stopping data into a single training dataset and created the final model used to predict on the validation set. Preliminary analysis indicated this technique moderately improved results. Early stopping monitored micro-F1 with a patience of 5. The weights of each LM were re-initialized between validation folds.

4.4 Statistical Significance

t-tests are a commonly used test of statistical significance (Demšar, 2006). They require a set of evaluation results which are typically collected over each round of cross-validation or over multiple datasets. Paired tests provide more statistical power for detecting differences than unpaired tests and therefore care should be taken to ensure test sets are identical for all systems during cross-validation. Briefly, a t-test assumes that the two methods perform equivalently (this is the null hypothesis), and any difference can be explained by random variation. To test this, the t-test compares two distributions and calculates a test statistic which is used to calculate a p-value which indicates the probability of accepting the null hypothesis. In this case, the probability that the two systems are equivalent. Therefore, low p-values indicate a low probability that the systems are equivalent, meaning they are statistically significantly different. α, a cut-off threshold for p is used to draw conclusions. When $p < \alpha$ one can conclude that the two systems are significantly different. $\alpha = 0.05$ is commonly used and indicates a 95% probability that the systems are different. Since we are testing for a difference in the better (greater than) or worse (less than) direction, a two-tailed test is used.

Although t-tests are commonly used, they are not well-suited for comparing results over cross-validation or over multiple datasets. This is because the t-test assumes that the scores are independent and normally distributed. The scores across cross-validation are not independent because the training sets of each round overlap and the results may not be normally distributed. When comparing across datasets the scores are independent, but are not normally distributed. Several corrections to t-tests have been proposed to account for the non-independence of results (Nadeau and Bengio, 2003; Grandvalet and Bengio, 2006; Santafe et al., 2015). These tests update the estimate of variance to account for the non-independence of samples. An alternative to the t-test is the Wilcoxon signed-rank test (Wilcoxon, 1945). This is a non-parametric test which does not assume the data is independent nor normally distributed. Like the t-test, Wilcoxon produces both a test statistic and a p-value which may be used to determine statistical significance. Because of the few assumptions of non-parametric tests, Wilcoxon tests are arguably preferable to t-tests, even corrected t-tests (Demšar, 2006; Rainio et al., 2024) for comparing across rounds of cross-validation and across datasets, however because they are non-parametric they require more samples than parametric tests to make sound statistical inference. Importantly, Wilcoxon cannot produce a p-value less than 0.05 with only 5 samples, meaning, in our study it is impossible to conclude any two systems are different when comparing across 5 rounds of cross-validation. Therefore, we use Grandvalet and Bengio's T-test for Cross-Validation (Grandvalet and Bengio, 2006) as described by Santafe et al. (Santafe et al., 2015) when comparing two systems across rounds of cross-validation and the Wilcoxon signed-rank test when two systems comparing across datasets. For the T-Test for Cross-Validation we use a $corr = 0.35$ because it is a middle value balancing over-estimating and under-estimating statistical significance. To compare multiple systems we use the Friedman test, which is a non-parametric test designed for comparing multiple systems. While it is possible to use repeated t-tests or Wilcoxon tests to determine differences between many systems, repeated tests increase the risk of error (Rainio et al., 2024; Demšar, 2006) across the whole family of tests (e.g. when $\alpha = 0.05$ there is a 95% chance of being correct, but with 2 tests there is a $95\% * 95\% = 90.25\%$ chance of being correct for all tests.

5 Results

Table 3 shows the micro-F1 performance averaged over 5-fold cross validation for all LMs and datasets. Each row indicates the results for a dataset and each column indicates the results for a LM. Bold values indicate the highest score for that dataset. Rows are grouped by dataset type; first sci-

Dataset	BERT Model								
	English	Scientific			Clinical			Social Media	
	BERT	Bi-B	PMed-B	Blue-B	BiClin-B	BiDis-B	BlueB+M	BT	BiRed-B
BC5CDR	0.9540	0.9576	0.9518	**0.9610**	0.9545	0.9551	0.9555	0.2049	0.9579
BC7DCPI	0.8255	0.8274	**0.8300**	0.8293	0.8256	0.8260	0.8254	0.1349	0.8233
NCBI	0.8790	0.8693	**0.8835**	0.8765	0.8727	0.8670	0.8470	0.2081	0.8727
NLMChem	0.9354	0.8527	**0.9691**	0.9392	0.8513	0.8517	0.9347	0.1395	0.8521
i2b2 2010	0.8848	0.8883	0.8861	0.8927	0.8895	0.8893	**0.8988**	0.1607	0.8885
n2c2 2018	0.9142	0.9176	**0.9224**	0.9141	0.9178	0.9196	0.9183	0.0290	0.9180
BC7Med	0.5844	0.5894	0.5865	0.6151	0.6096	0.6078	**0.6732**	0.2957	0.6151
COMETA	0.6111	0.6108	0.6092	**0.6127**	0.6077	0.6074	0.6065	0.1359	0.6122
ADEMiner	0.5353	0.5289	0.5085	0.5193	0.5113	0.5192	0.5269	0.3104	**0.5509**
Mean	0.7915	0.7824	0.7914	0.7955	0.7822	0.7826	**0.7985**	0.1799	0.7878

Table 3: Validation set micro-averaged F1 scores averaged over 5-fold cross-validation.

entific, then clinical, then social media. Columns are similarly grouped. The final row shows the average performance of each LM over all datasets. An important note is that these results are not necessarily directly comparable to other reported results, since these are the average micro-f1 of 5-fold cross validation, not results on the provided test set (if provided). From this table we can make several observations:

Observation 1: BT performs poorly on all datasets, even BC7Med and ADEMiner which consist of Tweets. Recall that BT was trained on general English Tweets and not on biomedical data. This implies that modeling biomedical language is more important than modeling the short-form structure of Twitter. Therefore, when choosing a LM it may be best to consider the domain first and the target task's structure second.

Observation 2: PMed-B performs the best for most datasets, but it doesn't have the highest average among all datasets. This is because it doesn't do as well as other LMs on social media datasets. We suspect this is related to its difference in vocabulary, but it may also be because it has never been trained for the general English domain which social media posts may most closely reflect. Furthermore, although PMed-B performs the best for most scientific datasets it is never statistically significantly better than Bi-B or Blue-B (the other scientific models), but it is statistically significantly worse than both Blue-B ($p = 0.0032$ and Bi-B ($p = 0.0099$) on BC5CDR.

Observation 3: The results for most datasets are similar no matter what LM you choose (except BT). For example, the scores for BC5CDR (excluding BT) range from 0.9518 to 0.9610, a difference of only 0.0092 = 0.92%. This pattern of similar performance is observed in nearly all datasets. For 4 out of the 9 datasets the difference between best and worst performing (excluding BT) is less than 1%, for 7 out of the 9, this difference is less than 5%. The only large differences are for NLMChem (11.78%) and BC7Med (8.88%). Furthermore, when comparing the average performance across all datasets (the Mean row), the difference between the best and worst performing (excluding BT) is only 1.63%. This observation is supported by statistical significance testing. Using Friedman's test to compare the performance of all LMs across all datasets reveals a statistically significant difference between the LMs ($p = 0.0017$), but when removing BT from the comparison, Friedman's test calculates a $p = 0.4061$ revealing no statistically significant difference between any system except BT when compared across all datasets.

Observation 4: BERT is the base model from which all LMs (except PMed-B) are adaptively pretrained. However, despite this, the change in performance is minimal (excluding BT) for most datasets; less than 2% for 7 out of 9 datasets, 3.37% for NLMChem, and 8.88% for BC7MED. However, when comparing one dataset at a time across cross-validation scores, we found that at least one (but never all) adaptively pretrained system significantly improved upon BERT. This was true for all datasets except for NCBI and ADEMiner. This indicates that although significant improvement is possible it is not guaranteed.

Observation 5: Common wisdom is to choose a LM most closely related to your target domain. However, the results in Table 3 show that P-MedB (a scientific LM) performs the best for n2c2 2018 (a clinical dataset). Similarly for the social media datasets of BC7Med and COMETA, the best performing systems are BlueB+M (a clinical LM) and PMed-B (a scientific LM) respectively. At first

glance this would imply that common wisdom is not always advisable, however comparing across folds of cross-validation for n2c2 2018 the p-value between P-MedB and BiDis-B (the best performing clinical model) was $p = 0.0686$, indicating no statistical significance between these systems. Similar results of $p = 0.5078$ and $p = 0.6720$ were found for BC7Med and COMETA respectively when comparing against BiRed-B (the best performing Social Media LM). Therefore, common wisdom stands. You should choose a LM that is most closely related to your domain.

Observation 6: Similar LMs perform similarly. Friedman's test from observation 3 revealed that no two LMs other than BT are statistically significantly different when compared across all datasets. Not surprisingly, using Friedman'a test for within category performance across within category datasets (e.g. clinical LMs - BiClin-B, BiDis-B, BlueB+M - across all clinical datasets - i2b2 2010, n2c2 2108) also showed no statistical significance. $p = 0.1690$ for biomedical systems on biomedical datasets, $p = 0.8948$ for clinical systems on clinical datasets. Using Wilcoxon within the social media datasets reveals a statistical significance $p = 0.0039$ due to BT's poor performance.

We also compared within-category systems across folds of cross-validation for each dataset using t-test for cross-validation and found little significant differences. There was a single significant difference between scientific models on scientific datasets which was on BC5CDR for which (as previously noted) PMed-B performed significantly worse than both Bi-B and Blue-B. Among clinical systems there were no statistically significant differences on n2c2 2018, but on i2b2 2010 we found that BlueB+M is statistically significantly better than BiClinB ($p = 0.0192$). Lastly, as expected for all social media datasets BiRedB is statistically significantly better than BT on all datasets (BC7Med $p = 0.0035$; COMETA $p = 0.0073$, ADEMiner $p =< 0.0001$).

6 Conclusions and Limitations

In this paper we performed a comparison of nine different LMs on nine biomedical datasets spanning the scientific, clinical, and social media subdomains on the task of NER. We confirmed that you should choose a LM most closely related to your target domain, that models trained on similar data will perform similarly, and that only moderate improvements are gained by adaptive pretraining. However, we were surprised by how little the differences between systems typically were, particularly with respect to BERT versus its adaptively pretrained counterparts, and most particularly by PMed-B's only slight improvement despite huge computational costs.

Based on our observations we can make a few suggestions:

1) As an NER developer in a specialized domain, follow common wisdom and select a model most closely related to your target task. It is unnecessary to generate and compare results across many different LMs. Their performance difference will very likely be statistically insignificant. This is a relief particularly because NER is often a basic task used within larger systems. It frees up time and effort to focus on other system components and other hyperparameter tuning.

2) As researcher in a novel domain for which no adaptively pretrained models exist, you can prototype using BERT. Performance gains by adaptively pretraining may improve results, but don't expect a dramatic change. If results are promising, then it may be worthwhile to adaptively pretrain a model, but if results are not promising, focus on developing a new technique or annotating more data.

3) Pretraining a model from scratch is not worth the costs for all but the most important or widely-used domains. It can improve performance but adaptively pretraining is a comparable method that is computationally more efficient.

We wonder if these results generalize to other lower-level tasks, such as relationship extraction or span detection, and if these results generalize to large language models (LLMs) such as LLaMa (Touvron et al., 2023). Since LLMs are recognized as being able to better generalize to new data, we feel this is likely and believe that these results suggest that using non-adapted foundation models may be sufficient. This is particularly salient since adapting pretrained LLMs is typically prohibitively expensive for most researchers and developers.

Lastly, we acknowledge limitations: (1) Significance tests were performed across a small number of samples which reduces their power of inference. Furthermore, since they compare across a distribution of F1 scores they don't take into account the number of samples in each dataset. (2) LM and dataset specific hyperparameter tuning could reveal more stark differences between LMs.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):1–14.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Rezarta Islamaj Dogan and Zhiyong Lu. 2012. An improved corpus of disease mentions in PubMed citations. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 91–99.

Kevin Donnelly and 1 others. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.

Graciela Gonzalez, Ari Z Klein, Ivan Flores, Davy Weissenbacher, Arjun Magge, Karen O'Connor, Abeed Sarker, Anne-Lyse Minard, Elena Tutubalina, Zulfat Miftahutdinov, and 1 others. 2020. Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*.

Yves Grandvalet and Yoshua Bengio. 2006. Hypothesis testing for cross-validation. *Montreal Universite de Montreal, Operationnelle DdIeR*, 1285.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, and 1 others. 2021. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(1):1–12.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2020. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. *arXiv preprint arXiv:2010.10391*.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Claude Nadeau and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific reports*, 14(1):6086.

Richard J Roberts. 2001. PubMed Central: The GenBank of the published literature.

Guzman Santafe, Iñaki Inza, and Jose A Lozano. 2015. Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, and 1 others. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, 2016.

Davy Weissenbacher, Siddharth Rawal, Arjun Magge, and Graciela Gonzalez-Hernandez. 2021. Addressing extreme imbalance for detecting medications mentioned in twitter user timelines. In *International Conference on Artificial Intelligence in Medicine*, pages 93–102. Springer.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30.

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, and 1 others. 2007. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Trade-offs in Medical LLM Adaptation: An Empirical Study in French QA

Ikram Belmadani[1,2] Oumaima El Khettari[2] Carlos Ramisch[1]
Frederic Bechet[1] Richard Dufour[2] Benoit Favre[1,3]

[1]Aix-Marseille Univ., CNRS, LIS UMR 7020, 13000 Marseille, France,
[2]Nantes Univ., École Centrale Nantes, CNRS, LS2N UMR 6004, 44000 Nantes, France,
[3]Grenoble Alpes Univ., CNRS, INRIA, Grenoble INP, LIG UMR 5217, 38000 Grenoble, France
Correspondence: first.last@univ-amu.fr, univ-nantes.fr

Abstract

The development of large language models (LLMs) has led to an increased focus on their adaptation to specialized domains and languages, yet the effectiveness of domain adaptation strategies remains unclear. We present a study of medical domain adaptation using French medical question-answering (QA) as a case study. We compare continual pretraining (CPT), supervised fine-tuning (SFT), and their combination across three model families, multiple sizes, and three initialization types, explicitly disentangling adaptation effects from base model choice. We evaluate both multiple-choice (MCQA) and open-ended QA (OEQA) under greedy and constrained decoding using automatic metrics and LLM-as-a-Judge evaluation. For MCQA, CPT+SFT most often achieves the best scores, but gains over SFT are small and frequently not statistically significant, making SFT a strong and cost-effective default. For OEQA, CPT consistently improves overlap-based metrics, while SFT often degrades generation quality; instruction tuning and CPT+SFT are preferred by LLM-based evaluation. Cross-lingual experiments further show effective transfer from French adaptation to English benchmarks. Overall, we provide practical guidelines for selecting adaptation strategies under computational constraints.

1 Introduction

LLMs are increasingly applied to medical question-answering (QA) and clinical reasoning, where accuracy, robustness, and domain-specific knowledge are critical (Huang et al., 2024). However, most high-performing LLMs are trained on general-domain data, making domain adaptation essential for safe medical deployment. In practice, this adaptation relies on continual pretraining (CPT) on domain corpora and supervised fine-tuning (SFT) on task-specific data.

Despite their widespread use, the relative effectiveness of these strategies remains unclear. Their impact depends on training scale, data composition, and optimization choices (Christophe et al., 2024; Lu et al., 2025), and even combined approaches yield inconsistent or statistically insignificant gains (Jeong et al., 2024a).

Most prior work fixes base model initialization, making it difficult to disentangle adaptation effects (Lu et al., 2025; Christophe et al., 2024). Evaluations are also predominantly English-centric, and largely focus on MCQA, limiting interpretation and generalization, especially given recent evidence of memorization in medical LLMs (Li et al., 2025).

More broadly, this work is motivated by a practical constraint often overlooked in the literature. In many real-world settings, especially for non-English medical NLP, both domain-specific data and computational resources are limited. As a result, the key question is not whether one strategy can theoretically outperform another, but how to allocate limited resources effectively. In this context, practitioners face a concrete and unresolved question: *given limited data and computational resources, which adaptation strategy should be prioritized?* Existing studies provide partial answers, but heterogeneous setups make it difficult to derive actionable guidelines.

To address these limitations, we conduct a controlled study of medical domain adaptation using French medical QA as a case study. We compare CPT, SFT, and their combination across model families and sizes while varying base initialization, and evaluate models in both French and English to isolate domain and cross-lingual effects.

We include OEQA as a complementary evaluation of generative behavior, but note that its assessment remains challenging; our conclusions are therefore primarily grounded in MCQA.

Our goal is to provide practical guidance on when and why CPT and SFT are effective under

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 209–234
July 3-4, 2026 ©2026 Association for Computational Linguistics

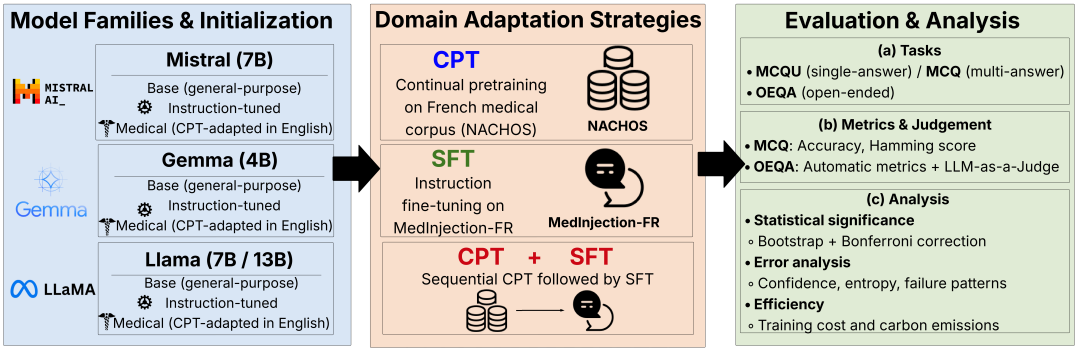


Figure 1: Overview of the experimental pipeline for evaluating medical domain adaptation strategies.

realistic constraints. Our study is guided by the following research questions:

- **RQ1:** *What are the performance and efficiency trade-offs between CPT and SFT across model families and sizes?*
- **RQ2:** *How does base model initialization influence the effectiveness of CPT and SFT for medical domain adaptation?*
- **RQ3:** *How does French medical adaptation affect cross-lingual transfer to English?*

Our contributions are: (i) we introduce a controlled and reproducible framework to compare medical domain adaptation strategies across model families, sizes, initialization types, and decoding settings; (ii) we provide a statistically grounded analysis of CPT and SFT for medical QA, covering performance trade-offs, error patterns, and cross-lingual transfer to English benchmarks. All resources are publicly available: `https://github.com/ikram28/MedAdapt`.

2 Related Work

Medical LLM adaptation primarily relies on CPT on domain-specific corpora and SFT on instruction–response data, both shown to support domain transfer (Gururangan et al., 2020; Gema et al., 2024). CPT has been adopted in models such as MediTron (Chen et al., 2023b), BioMistral (Labrak et al., 2024a), PMC-Llama (Wu et al., 2023), and MedGemma (Sellergren et al., 2025). However, recent analyses question the robustness and consistency of CPT gains under stricter evaluation protocols (Jeong et al., 2024a). In parallel, SFT-based models such as ChatDoctor (Li et al., 2023) and MedAlpaca (Han et al., 2023) report great task-level improvements, though evaluations remain largely in English.

Medical domain adaptation is further challenged in non-English settings due to limited domain-specific resources. Several multilingual medical LLMs have been proposed, including Medical mT5 (García-Ferrero et al., 2024), BiMediX (Pieri et al., 2024), Apollo (Wang et al., 2024a), and MMedLM (Qiu et al., 2024). However, these models are mostly evaluated on translated benchmarks, with limited validation on native-language medical tasks, leaving their language- and cultural-specificities underexplored. Evaluation practices also pose challenges. Widely used benchmarks such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022) primarily target English.

Beyond proposing individual models, recent work has compared adaptation strategies in controlled settings. Christophe et al. (2024) analyze CPT, SFT, and related techniques for clinical LLMs, finding that CPT alone yields limited gains but can amplify performance when combined with instruction tuning. Similarly, Lu et al. (2025) study CPT, SFT, and preference-based optimization across domains, highlighting complex interactions between adaptation methods. However, these studies focus on English and fix the base model initialization. In contrast, in this work, we systematically compare CPT, SFT, and their combination across multiple model families and initialization points for French medical QA, while also evaluating cross-lingual performance and analyzing adaptation behavior.

3 Experimental Framework

We propose a controlled experimental framework to evaluate medical domain adaptation strategies across architectures, initialization points, and task formats, as illustrated in Figure 1. Our setup explic-

itly varies (i) the base model and its prior training, (ii) the adaptation strategy, and (iii) the evaluation task and language in order to isolate the factors that drive adaptation effectiveness.

3.1 Base Models and Adaptation Approaches

Our study focuses on model families with three complementary initialization states: (i) a general-purpose base model, (ii) an instruction-tuned variant, and (iii) a medically adapted version obtained via CPT. This constraint is central to our experimental design, as it enables a controlled comparison that isolates the effect of adaptation strategy from that of the starting point. As a result, model selection is restricted to families providing these aligned variants, rather than to the most recent model releases.

We consider three model families spanning different sizes, pretraining regimes, and linguistic exposure. Specifically, we include Mistral-7B, Gemma-4B, and Llama models at the 7B and 13B scales. For Mistral-7B, we use Mistral-7B-v0.1 and its instruction-tuned version, and BioMistral-7B, a model adapted to the biomedical domain via CPT (Jiang et al., 2023; Labrak et al., 2024b). For Gemma, we rely on the Gemma-3-4B pretrained and instruction-tuned models, together with MedGemma-3-4B, which incorporates medical pretraining (Team et al., 2025; Sellergren et al., 2025). Finally, for Llama, we include both 7B and 13B variants, using the base and chat versions of Llama-2, as well as their medically adapted counterparts, MediTron-7B and MedLlama-13B (Touvron et al., 2023; Chen et al., 2023a; Wu et al., 2024).

These families differ not only in scale but also in pretraining data and exposure to French. Mistral and Gemma are explicitly multilingual, whereas Llama models are primarily English-centric, although exact language proportions are not disclosed. Except for MedGemma, whose medical pretraining corpus is not fully documented, all medical variants rely on PubMed Central as their primary biomedical source[1].

Across all model families and initialization points, we investigate three adaptation strategies: (i) CPT on domain-specific corpora, (ii) SFT on instruction-response pairs, and (iii) a sequential CPT+SFT pipeline.

[1] https://pmc.ncbi.nlm.nih.gov/

3.2 Training Data

CPT. We use NACHOS corpus (Labrak et al., 2023), an open-source French medical dataset comprising 4 GB of text collected from French medical websites; full details are provided in Appendix A.

SFT. We use the train and validation sets of the MedInjection-FR corpus (Belmadani et al., 2026b), which contains 543 505 instruction-response pairs. The dataset includes multiple-choice questions with a single **unique** correct answer (MCQU, ∼83%), **multiple** correct answers (MCQ, ∼6%), and OEQAs (∼11%). This mixture allows us to evaluate adaptation effects across both discriminative and generative medical reasoning tasks. Additional dataset details are provided in Appendix B.

3.3 Training Process

To explore the trade-off between computational cost and model plasticity, we adopt contrasting fine-tuning regimes for CPT and SFT. CPT is performed using full-parameter fine-tuning, while SFT relies on parameter-efficient adaptation. This choice is supported by preliminary experiments, as explained in Appendix E.

CPT. CPT is performed for three epochs following the setup of Labrak et al. (2024b). Full hyperparameter details are provided in Appendix C.

SFT. We employ DoRA (Weight-Decomposed Low-Rank Adaptation) (Mao et al., 2024), an extension of LoRA (Hu et al., 2022) that decouples magnitude and directional updates. We select DoRA after preliminary experiments, as detailed in Appendix E. SFT is run for ten epochs, with hyperparameters reported in Appendix D.

3.4 Evaluation Protocol

Benchmarks. We evaluate all models on MedInjection-FR test set, which consists of 14 533 native French medical examples and 13 293 translated examples derived from established English benchmarks. The test set covers MCQU, MCQ, and OEQA tasks, enabling evaluation of both answer selection and free-form answers. Benchmark sources and translation procedure are detailed in Appendix G.

Prompting Strategy. All evaluations are conducted in a zero-shot setting using greedy, deterministic decoding to ensure reproducibility. For MCQU and MCQ tasks, following Liang et al.

(2022); Beeching et al. (2023); Chen et al. (2023a), we restrict the output vocabulary to valid answer options to prevent hallucinated responses. To mitigate position bias, we randomly shuffle answer choices three times and report aggregated results, following best practices for MCQ evaluation (Pezeshkpour and Hruschka, 2024). Prompt templates are provided in Appendix H.

Evaluation Metrics. For MCQU, we report *Exact Match (EM)*, which measures the proportion of questions for which the predicted answer exactly matches the gold answer. For MCQ, we additionally report the *Hamming score*, which accounts for partial overlap between predicted and reference answer sets and is therefore more informative for multi-answer questions. Formal definitions of both metrics are provided in Appendix F.

For OEQA, we rely on both automatic text-based metrics and model-based judgments. We report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2019) as automatic baselines. Their reliability was assessed through agreement with senior physician annotations on a held-out subset of 500 OEQA instances in Belmadani et al. (2026a), where MedGemma-27B was identified as the most stable and best-performing LLM judge, and is therefore used in the present work.

Statistical Significance and Error Analysis. We assess statistical significance using a percentile bootstrap procedure with 10 000 resamples, following Jeong et al. (2024b). Differences between paired model configurations are considered significant when the associated two-sided p-value is below a predefined threshold α. To control for multiple comparisons, we apply the Bonferroni correction, yielding a corrected α as specified in Appendix J. In addition, we conduct an error analysis by examining output probabilities, confidence scores, and entropy, enabling us to characterize how CPT and SFT affect uncertainty and error patterns across different base model initializations.

4 Results and Discussion

4.1 MCQA Evaluation

Table 1 reports performance on MCQA across three model families (Gemma-4B, Mistral-7B, Llama-7B-13B), three initialization types (General, Instruct, Medical), and three adaptation strategies (CPT, SFT, CPT+SFT). Results are shown for both MCQs and MCQUs, using EM and Hamming scores for MCQs. All results reported in this table are obtained using constrained decoding. The corresponding results under greedy decoding are provided in Appendix I.

Effectiveness of Adaptation Strategy: A recurring pattern observed throughout the results is:

$$\text{BASE} \ll \text{CPT} < \text{SFT} \lesssim \text{CPT+SFT}$$

The strongest performance is most frequently achieved by the CPT+SFT adaptation. Across model families and initialization types, CPT+SFT yields the highest scores in aggregated EM as well as in MCQ and MCQU EM more often than any other strategy.

However, a closer inspection of the results indicates that the gains brought by CPT+SFT over SFT alone are generally limited. When CPT+SFT attains the highest score, the margin over SFT rarely exceeds 1.3 points. In contrast, in configurations where SFT outperforms CPT+SFT, the performance gap is larger. For example, on Llama-7B Instruct, SFT exceeds CPT+SFT by 3.12 points, and a similar pattern is observed for Mistral-7B Instruct, with a gap of 1.44 points in favor of SFT.

Furthermore, the statistical analysis reported in Appendix J shows that, when comparing each adapted model to its corresponding base model, the observed improvements are not always statistically significant. In particular, for Gemma Instruct, neither SFT nor CPT+SFT yields statistically significant gains over the base model. Likewise, for Llama-7B Instruct, the improvement brought by CPT+SFT is not statistically significant. These constitute the only cases in which the improvements of SFT or CPT+SFT over the base model fail to reach statistical significance. Consequently, although CPT+SFT most frequently ranks first, its advantage over SFT is not consistently substantial.

By contrast, CPT alone exhibits less stable behavior. Although it can improve performance in some rare cases, it can also occasionally degrade performance compared to the base model. Moreover, it is the strategy that most often fails to produce statistically significant improvements over the base. This is the case for 8 models: MedGemma, all Llama-13B variants, Llama-7B GENERAL and INSTRUCT, and Mistral-7B GENERAL and MEDICAL. This suggests that representation-level domain adaptation is most effective when paired with task-specific supervision.

Model Type	Strategy	MCQ		MCQU	Aggregation	OEQA		
		EM	Hamming	EM	EM	ROUGE-L	BERT-F1	Judge
Gemma-4B								
GENERAL	Base	2.24	30.17	27.11	14.68	7.11	46.62	25.09
	CPT	0.73	11.71	25.83	13.28	**10.18**	**49.07**	**25.71**
	SFT	3.73	**43.57**	32.36	18.05	8.23	47.76	21.60
	CPT+SFT	**3.90**	42.81	**32.59**	**18.25**	6.01	45.78	24.80
INSTRUCT	Base	**4.83**	44.01	29.30	17.06	**7.38**	**48.94**	**47.71**
	CPT	3.47	46.00	25.05	14.26	4.57	42.77	13.35
	SFT	3.68	**48.22**	**31.95**	**17.81**	2.20	38.26	20.21
	CPT+SFT	3.42	48.14	30.73	17.07	3.12	40.90	20.76
MEDICAL	Base	1.98	31.46	26.41	14.19	**7.38**	**48.94**	**22.01**
	CPT	1.68	24.04	25.13	13.41	6.77	45.18	1.95
	SFT	**3.54**	**46.22**	30.66	17.10	5.70	45.66	14.23
	CPT+SFT	3.38	43.28	**30.86**	**17.12**	5.04	43.05	11.41
Mistral-7B								
GENERAL	Base	0.37	5.40	28.52	14.44	5.82	44.21	**27.23**
	CPT	3.54	30.50	27.21	15.37	7.22	46.32	24.57
	SFT	5.24	21.62	**32.88**	19.06	**8.83**	**48.02**	22.93
	CPT+SFT	**6.13**	**30.86**	32.29	**19.21**	6.62	46.35	24.89
INSTRUCT	Base	4.86	23.53	24.92	14.89	7.34	49.66	30.14
	CPT	**7.32**	**36.18**	28.79	18.06	**13.51**	**53.87**	**37.59**
	SFT	6.80	23.42	**31.61**	**19.21**	12.41	52.72	17.63
	CPT+SFT	5.45	32.47	30.09	17.77	9.02	48.99	32.13
MEDICAL	Base	2.80	17.47	26.69	14.74	11.34	**51.58**	20.76
	CPT	3.57	24.43	25.73	14.65	**12.41**	51.45	17.89
	SFT	3.36	26.37	31.62	17.49	8.75	47.86	16.72
	CPT+SFT	**4.94**	**27.27**	**32.58**	**18.76**	9.15	48.63	**24.96**
Llama-7B								
GENERAL	Base	1.33	12.01	25.72	13.53	5.05	41.27	9.39
	CPT	1.12	12.86	25.59	13.36	**10.58**	**47.85**	3.78
	SFT	2.66	28.41	28.93	15.80	6.02	44.49	7.67
	CPT+SFT	**3.17**	**46.00**	**29.89**	**16.53**	5.85	44.67	**12.26**
INSTRUCT	Base	3.95	34.43	25.08	14.51	2.57	43.85	25.35
	CPT	3.93	**42.67**	25.07	14.50	11.16	**51.37**	26.06
	SFT	**5.12**	21.06	**29.32**	**17.22**	**11.44**	51.28	12.92
	CPT+SFT	3.13	25.98	25.07	14.10	9.92	50.99	**27.84**
MEDICAL	Base	0.23	2.90	24.43	12.33	5.61	43.25	12.50
	CPT	2.37	28.06	25.60	13.99	**8.00**	**45.79**	13.14
	SFT	3.24	29.10	30.44	16.84	5.40	42.25	9.50
	CPT+SFT	**3.80**	**44.95**	**31.52**	**17.66**	5.87	44.34	**17.39**
Llama-13B								
GENERAL	Base	2.14	21.20	26.11	14.13	2.10	33.25	11.79
	CPT	2.53	19.17	26.99	14.76	**14.12**	**50.36**	5.66
	SFT	**3.54**	**40.49**	30.95	17.24	5.60	43.19	14.88
	CPT+SFT	3.34	29.59	**32.36**	**17.85**	6.30	45.45	**20.38**
INSTRUCT	Base	0.09	29.74	21.52	10.81	3.40	45.31	30.02
	CPT	5.63	**37.40**	25.01	15.32	12.34	**53.07**	**36.19**
	SFT	6.58	23.96	30.20	18.39	11.54	50.94	11.81
	CPT+SFT	**7.77**	25.26	**31.58**	**19.68**	**12.86**	52.46	20.22
MEDICAL	Base	1.77	11.82	24.62	13.19	5.00	42.11	10.86
	CPT	2.26	30.87	24.10	13.18	8.00	45.79	13.39
	SFT	3.12	41.24	30.62	16.87	6.85	45.22	13.77
	CPT+SFT	**3.24**	**45.59**	**32.25**	**17.74**	**8.38**	**46.29**	**19.55**

Table 1: Constrained decoding results (%) for MCQ/MCQU and OEQA. Aggregation corresponds to average EM over MCQ and MCQU. **Bold** denotes the best strategy, and underlining the best initialization.

Overall, while CPT+SFT ranks first most often, its limited and inconsistent gains over SFT, together with a substantially higher computational cost (see Appendix O), make SFT a strong default for medical MCQA. For example, on 7B models, CPT+SFT costs over $1 500 versus $360 for SFT, with a fourfold increase in carbon emissions.

Impact of Model Initialization: The impact of model initialization (General / Instruct / Medical) varies across MCQA metrics and question formats. Considering the overall best scores across all model families, instruction-tuned models dominate the most demanding EM settings: the highest MCQ EM and aggregated MCQA EM scores are both achieved by Llama-13B Instruct, while the best MCQ Hamming score is obtained by Gemma-4B Instruct. In contrast, the best MCQU EM score is achieved by a general Mistral model.

At the family level, the patterns differ. For MCQ EM, the best score within each model family is always obtained by an instruction-tuned variant, confirming that instruction alignment is particularly beneficial for exact multi-label prediction; this ad-

vantage is further supported by statistically significant gains when compared to general or medical initializations (see Appendix J). For MCQ Hamming, results are more balanced, with the best scores split across initialization types (two instruction-tuned, one general, and one medical).

For MCQU EM, general models most frequently achieve the best performance (three cases), followed by medical models, while instruction-tuned models do not dominate. This indicates that when only a single answer must be selected, performance is driven primarily by answer plausibility ranking, favoring strong language modeling and domain knowledge, while explicit instruction alignment, which mainly benefits structured or multi-label outputs, provides less advantage. Finally, for the aggregated MCQA score, no single initialization consistently dominates: instruction-tuned and general models each obtain the best result in two configurations (with ties between them), while medical models lead in one case, and differences across initializations are often not statistically significant.

4.2 OEQA Evaluation

The right side of Table 1 reports OEQA across model families using ROUGE-L, BERTScore, and LLM-as-a-Judge. We additionally report BLEU and METEOR in Appendix I, as they reflect similar information to ROUGE-L. Overall, absolute scores remain moderate, reflecting the difficulty of evaluating free-form answer generation. ROUGE-L scores should be interpreted with caution, as they measure surface-level lexical overlap and penalize semantically correct answers that differ in formulation (Yim et al., 2025; Zhu et al., 2025).

Moreover, OEQA represents only 11% of the training data, resulting in a strongly imbalanced supervision signal. Models are therefore adapted to generate short, structured outputs (answer letters in MCQA), which limits OEQA performance.

Effect of Adaptation Strategy: Across model families, SFT often degrades ROUGE-L and BERTScore-F1 compared to base or CPT-adapted models, particularly for instruction-tuned and medical variants. This suggests that SFT can overly constrain generation, reducing lexical diversity and semantic overlap in an open-ended setting.

By contrast, CPT is the most consistently beneficial strategy for OEQA. CPT improves ROUGE-L and BERTScore-F1 across most general, instruct, and medical models, with especially strong gains for Mistral and Llama families. These results suggest that domain-adaptive language modeling supports better medical generation than instruction-level supervision alone. Combining CPT with SFT rarely outperforms CPT alone and often leads to intermediate or degraded performance, reflecting the same instability observed in MCQA, but with more pronounced negative effects in OEQA.

In contrast to overlap-based metrics, LLM-as-a-Judge favors CPT+SFT in half of the configurations (6/12), compared to three cases each for the base and CPT models. Gains are most pronounced for Llama-7B, where CPT+SFT consistently outperforms SFT across initializations, and for medical models, where it yields the best or near-best qualitative scores. However, despite these trends, statistically significant improvements over the base model are rare: CPT is significant in only three cases, SFT in two (all involving smaller 4B models), and CPT+SFT never yields statistically significant gains over the base model in OEQA.

Effect of Model Initialization. Initialization effects on OEQA depend strongly on the evaluation metric. For overlap-based metrics, no initialization consistently dominates: ROUGE-L is split between general and instruction-tuned models, while medical models never achieve the top score; BERTScore-F1 is mostly dominated by instruction-tuned models, with a single exception (Gemma).

LLM-as-a-Judge reveals clearer and statistically grounded patterns. When differences are significant, medical models are consistently outperformed across families, particularly under SFT and CPT (Table 9). Comparisons between instruction-tuned and general models are mixed and direction-dependent, with some significant gains under SFT and CPT (notably for Mistral and Llama), but these effects largely disappear under CPT+SFT.

Overall, medical initialization alone does not improve OEQA, while instruction-tuned initialization yields more reliable, yet limited, gains when significant.

5 Cross-Lingual Transfer After French Medical Adaptation

To analyze whether models perform better in English prior to adaptation and how cross-lingual adaptation affects performance, we compute the EM accuracy difference for MCQU benchmarks as the score on French translations minus the score on the corresponding native English datasets. Figure 2

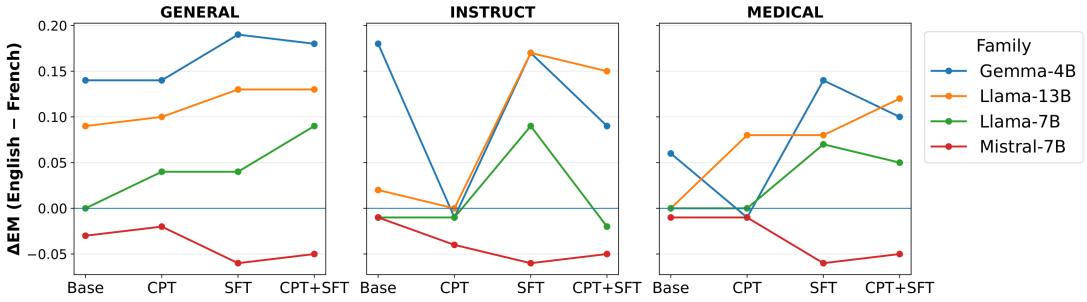


Figure 2: Difference in EM accuracy (ΔEM) between native English MCQU test benchmarks and their French translations across model families and adaptation strategies (constrained decoding).

reports averaged results across datasets using constrained decoding; full results for both greedy and constrained decoding are provided in Appendix M.

For the Mistral family, base models consistently perform better on the translated French benchmarks than on the original English data. This trend persists after adaptation with CPT, SFT, and CPT+SFT, with French performance systematically exceeding English, the differences being statistically significant (Table 13).

In contrast, Gemma and Llama families show higher performance on native English benchmarks at the base level, and this advantage remains after adaptation on French data. Moreover, adaptation gains are often larger in English than in French (Table 12), despite all adaptation data being in French.

These results suggest that Mistral models encode French more effectively, whereas Gemma and Llama have stronger English representations. Notably, the improvements observed in both languages indicate effective cross-lingual transfer of medical knowledge: adapting with French medical data improves performance on the original English benchmarks, sometimes more than on their French translations. This supports the complementarity of multilingual medical data, in line with Wang et al. (2024a).

A salient exception is Llama-7B: before adaptation, the base model shows slightly higher performance on French translations than on English, but this difference is not statistically significant (Table 13). After adaptation, English performance surpasses French, suggesting that adaptation amplifies the model's dominant English representations.

6 Effect of Translated Benchmarks on Performance and Confidence

We compare model behavior on a native benchmark, MediQAl (Bazoge, 2025), and a translated

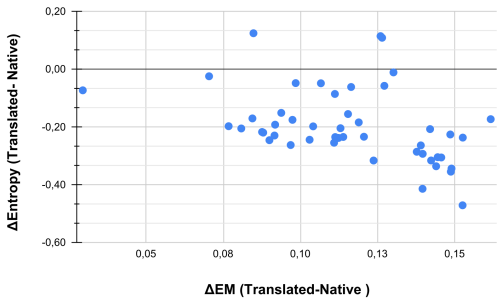


Figure 3: Relationship between accuracy gain (ΔEM) and change in predictive entropy ($\Delta Entropy$) when moving from the translated to native benchmarks. Each point corresponds to a model configuration.

benchmark, MedMCQA (Pal et al., 2022), using accuracy and confidence-based metrics. Both benchmarks consist of MCQUs of comparable size, for fair comparison. Although instances are not shared, consistent differences across models are observed.

As shown in Figure 3, all models achieve higher EM accuracy on the translated benchmark. This gain is systematically accompanied by a reduction in predictive entropy, indicating that translated benchmarks induce more confident and less uncertain predictions. The concentration of models in the bottom-right quadrant suggests that translated benchmarks operate in a different evaluation regime, characterized by both higher performance and reduced uncertainty.

Figure 5 further reveals that accuracy gains are often associated with increased confidence in incorrect predictions. Most models exhibit positive shifts in confidence even when wrong, indicating a systematic overconfidence effect induced by the translated benchmark.

Overall, these results show that translated benchmarks are not neutral substitutes for native ones: they tend to inflate performance while also altering

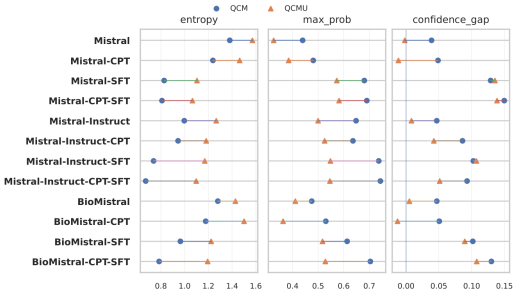


Figure 4: Probability-level metrics for MCQ and MCQU across Mistral variants.

model confidence calibration, potentially leading to over-optimistic evaluations.

7 Error Analysis

7.1 Probability-Level Analysis of MCQA

To explain why MCQ is harder than MCQU, we analyze class probability distributions from the Mistral family, selected for its high variance across models and adaptation settings. For each item, we compute entropy, maximum probability, and a confidence gap measuring gold/non-gold separation (mean gold vs. non-gold probability in MCQ; margin to the second-best option in MCQU). We also report a near-miss rate, defined as cases where all gold answers are ranked in the top-k but the predicted set is incorrect (Figure 4, Appendix K).

Across all variants, MCQ predictions are not more uncertain than MCQU: MCQ exhibits lower entropy and higher maximum probability, indicating confident local rankings. The confidence gap is consistently positive and increases with adaptation, but remains insufficient for exact multi-label generation under greedy decoding, leading to omissions or over-generation.

Adaptation clarifies this effect: SFT strongly improves MCQU, while gains on MCQ remain limited. CPT+SFT primarily increases ranking confidence rather than exact set match, yielding larger confidence gaps without reducing near-miss rates.

7.2 Verbosity Bias in OEQA

To better understand the differences observed between overlap-based metrics and LLM-as-a-Judge evaluations in OEQA, we analyze the length of generated answers across models. The results are reported in Appendix L. We find that CPT-adapted models systematically produce longer responses, with higher mean and median word counts across all model families. This increased verbosity provides a plausible explanation for their strong performance on ROUGE-L and BERTScore-F1, which reward lexical recall and content coverage.

In contrast, instruction-tuned models generate substantially shorter and more controlled answers, particularly under SFT, often producing concise responses with low variance. While this behavior negatively impacts overlap-based metrics, it aligns with higher LLM-as-a-Judge scores, suggesting that concise answers are preferred under LLM evaluation. Finally, SFT exhibits unstable behavior in OEQA, leading either to excessively short outputs or overly long responses depending on model initialization. Overall, these results indicate that OEQA performance is strongly influenced by length biases, and that improvements in automatic metrics may partially reflect increased verbosity rather than improved answer quality.

8 Conclusion

We presented a controlled and statistically grounded study of medical domain adaptation for LLMs using French medical QA, isolating the effects of model initialization, adaptation strategy, decoding, and evaluation. Our results show that adaptation effectiveness is task-dependent and that stronger strategies are not always more cost-effective. We therefore distill practical guidelines for selecting adaptation strategies based on data availability and computational constraints. Given the limited reliability of current OEQA metrics and the small proportion of OEQA supervision, our recommendations primarily emphasize MCQA, with OEQA trends interpreted cautiously.

Unlabeled data only. When only unlabeled medical text is available, CPT yields modest and unstable gains for MCQA and should not be used in isolation. Its benefits mainly appear on OEQA overlap-based metrics, which are sensitive to verbosity and should be interpreted with caution.

Labeled data only. With labeled QA data, SFT provides the best performance–efficiency trade-off for MCQA across all model families. It frequently matches or exceeds CPT+SFT while requiring substantially fewer computational resources, making it the most practical default in this setting.

Labeled and unlabeled data. When both data types are available, CPT+SFT most often achieves the highest MCQA scores, but improvements over

SFT are typically small and not consistently statistically significant. Consequently, CPT+SFT is justified only when maximal performance outweighs computational cost.

Initialization and compute considerations. Instruction-tuned models constitute the strongest baseline for French medical MCQA. Medical initialization alone does not reliably improve downstream performance. From a resource perspective, parameter-efficient SFT is by far the most cost-effective strategy, whereas CPT incurs high computational and environmental costs for limited MCQA gains, and CPT+SFT compounds these costs for marginal improvements.

Evaluation and transfer considerations. Finally, we observe strong evaluation effects: adaptation on French medical data transfers to English benchmarks, translated datasets inflate both accuracy and confidence, and OEQA metrics are sensitive to verbosity. These findings highlight the need for task-aware adaptation choices and cautious metric interpretation in medical LLM evaluation.

9 Limitations

Our evaluation of adaptation strategies faces several limitations. First, we perform an exploratory contamination study to assess possible exposure to NACHOS during pretraining (Appendix P). Although no direct evidence of memorization is observed, likelihood-based tests remain inconclusive due to the lack of a reliable non-member biomedical control corpus, requiring the use of synthetic controls. We therefore treat these results as indicative only and avoid causal conclusions about pretraining inclusion.

Second, our evaluation of OEQA relies on overlap-based metrics, BERTScore, and LLM-as-a-Judge. While these measures capture complementary aspects of answer quality, they do not fully characterize semantic equivalence, clinical correctness, or reasoning validity, and may therefore overlook qualitative differences between correct answers (Yim et al., 2025; Zhu et al., 2025).

Third, while we demonstrate the efficiency of SFT compared to CPT in terms of computational resources, our analysis does not account for the human effort required to create high-quality instruction-tuning datasets. This consideration is particularly relevant for low-resource settings where creating domain-specific instruction data may be costly.

Fourth, we do not include few-shot prompting as an evaluation setting. Our objective is to isolate the effects of parameter-level adaptation strategies under controlled and reproducible conditions. Few-shot prompting introduces additional sources of variance related to example selection, ordering, and prompt design, which would complicate statistical comparison and obscure the interpretation of adaptation gains. Moreover, few-shot prompting assumes access to curated task-specific examples at inference time, which may be unrealistic in medical deployment scenarios. For these reasons, we focus on zero-shot evaluation to ensure fair and stable comparisons across adaptation strategies.

Finally, our study focuses exclusively on CPT and SFT. We do not explore reinforcement learning–based adaptation strategies, such as preference optimization or reward-driven fine-tuning, which may better align models with clinical judgment or evaluation criteria. Investigating how such methods interact with CPT and SFT, particularly under multilingual and domain-specific constraints, constitutes an important direction for future work.

In addition, our findings about the effectiveness of adaptation strategies are specific to the medical domain and French language. The generalizability of these results to other domains or languages, particularly those with different resource constraints or linguistic characteristics, requires further investigation.

10 Ethical Considerations

This work is intended for research purposes only and not for direct clinical use. All experiments rely on publicly available biomedical datasets without identifiable patient data. We acknowledge that medical LLMs may generate inaccurate or overconfident outputs, particularly in open-ended settings, and therefore include physician-based evaluation protocols. We also report computational cost and estimated carbon emissions for transparency.

11 Acknowledgements

This work was financially supported by ANR MALADES (ANR-23-IAS1-0005). It was provided with computing HPC and storage resources by GENCI at IDRIS thanks to the grants 2025-AD011015256R1 and 2025-AD011016540 on the supercomputer Jean Zay's H100 partition.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adrien Bazoge. 2025. Mediqal: A french medical question answering dataset for knowledge and reasoning evaluation. *arXiv preprint arXiv:2507.20917*.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard hugging face. *Récupérée mai*, 24:2024.

Ikram Belmadani, Oumaima El Khettari, Pacôme Constant dit Beaufils, Richard Dufour, and Benoit Favre. 2026a. Who judges the judge? evaluating LLM-as-a-judge for French medical open-ended QA. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 142–157, Rabat, Morocco. Association for Computational Linguistics.

Ikram Belmadani, Oumaima el Khettari, Pacome Constant Dit Beaufils, Benoit Favre, and Richard Dufour. 2026b. Medinjection-fr: Exploring the role of native, synthetic, and translated data in biomedical instruction tuning. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2525–2544, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023a. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clement Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco AF Pimentel, and Shadab Khan. 2024. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10549–10561, Miami, Florida, USA. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mt5: An open-source multilingual text-to-text llm for the medical domain. *Preprint*, arXiv:2404.07613.

Aryo Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of LLaMA for the clinical domain. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 91–104, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024a. Medical adaptation of large language and vision-language models: Are we

making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Daniel P. Jeong, Pranav Mani, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. 2024b. The limited impact of medical adaptation of large language and vision-language models. *Preprint*, arXiv:2411.08870.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Zakaria Kaddari and Toumi Bouchentouf. 2022. Frbmedqa: the first french biomedical question answering dataset. *IAES International Journal of Artificial Intelligence*, 11(4):1588.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. Vera: Vector-based random matrix adaptation. *Preprint*, arXiv:2310.11454.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *LOUHI 2022*, Abou Dhabi, United Arab Emirates.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *Preprint*, arXiv:2304.00958.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024a. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024b. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Anran Li, Lingfei Qian, Mengmeng Du, Yu Yin, Yan Hu, Zihao Sun, Yihang Fu, Erica Stutz, Xuguang Ai, Qianqian Xie, and 1 others. 2025. Memorization in large language models in medicine: Prevalence, characteristics, and implications. *arXiv preprint arXiv:2509.08604*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wei Lu, Rachel K Luu, and Markus J Buehler. 2025. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(1):84.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A real-world medical Q&A benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand. Association for Computational Linguistics.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11662–11675, Bangkok, Thailand. Association for Computational Linguistics.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 16984–17002. Association for Computational Linguistics.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Preprint*, arXiv:2402.13963.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. A comprehensive survey of contamination detection methods in large language models. *arXiv preprint arXiv:2404.00699*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024a. Apollo: An lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. *arXiv preprint arXiv:2509.12405*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, Miami, Florida, USA. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A CPT Training Corpus : NACHOS Description

The NACHOS corpus is a French medical open-source dataset compiled through extensive web crawling and text collection. While the full corpus spans 7.4 GB of data and contains over one billion words sourced from 24 French-speaking high-quality websites (Labrak et al., 2023), we use in this work its *small* variant, NACHOS$_{small}$. This version consists of approximately 4 GB of data and was obtained by shuffling the full corpus and randomly selecting 25.3 million sentences to ensure homogeneous coverage of data sources.

Note: Full details of the corpus compilation and processing are available in the original paper (Labrak et al., 2023).

A.1 Corpus Composition

The NACHOS corpus encompasses a diverse range of medical textual sources, including:

- Descriptions of diseases and conditions
- Treatment and medication information
- General health-related advice
- Official scientific meeting reports
- Anonymized clinical cases
- Scientific literature
- Theses
- French translation pairs
- University health courses

A.2 Data Sources

The corpus integrates data from multiple sources, with the most significant contributions coming from:

- HAL (638,508,261 words)
- Haute Autorité de Santé (HAS) (113,394,539 words)
- Drug leaflets (74,770,229 words)
- Medical Websites Scraping (60,561,495 words)
- ANSES SAISINE (51,372,932 words)
- Public Drug Database (BDPM) (48,302,695 words)

A.3 Corpus Preparation

The researchers employed several preprocessing steps:

1. Text collection through web scraping, raw textual sources, and optical character recognition (OCR)
2. Sentence splitting using heuristic methods
3. Aggressive filtering to remove short or low-quality sentences
4. Language classification using a custom classifier trained on multilingual corpora

B SFT Training Corpus : MedInjection-Fr Description

B.1 Overview

MedInjection-FR (Belmadani et al., 2026b) is a large-scale French biomedical instruction dataset composed of native, translated, and synthetic instruction–response pairs. The dataset comprises 571 436 examples spanning MCQUs, MCQs, and OEQAs.

B.2 Data Composition

The dataset consists of 77 247 native examples, 417 674 translated examples, and 76 506 synthetic examples. All data are formatted as instruction–response pairs and normalized to a unified schema, ensuring consistency across heterogeneous sources and supervision types.

B.3 Quality Control for Translated Data

The translated subset was obtained by translating English biomedical instruction datasets into French using two LLMs: `GPT-4o-mini` (Hurst et al., 2024) and `Gemini 2.0 Flash` [2]. Translation quality was evaluated on the WMT 2024 Biomedical Translation Task benchmark (Neves et al., 2024) using BLEU and COMET metrics. `GPT-4o-mini` achieved a BLEU score of 51.01 and a COMET score of 0.8751, while `Gemini 2.0 Flash` achieved a BLEU score of 53.72 and a COMET score of 0.8783. These results are comparable to the best-performing system reported in the shared task (BLEU 53.54, COMET 0.8760), suggesting high semantic fidelity and robust preservation of biomedical terminology in the translated subset.

[2] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

B.4 Quality Control for Synthetic Data

The synthetic subset was generated using `GPT-4o` from source documents including clinical cases and biomedical abstracts. Each source document was used to generate multiple instructional tasks covering a broad range of biomedical reasoning, such as clinical summarization, factual QA, diagnostic reasoning, treatment suggestion, and classification.

To control generation quality, each synthetic instruction–response pair was evaluated using four independent large language models acting as automatic judges: `GPT-4.1-mini` [3], `Gemini 2.0 Flash`, `MedGemma-27B` (Sellergren et al., 2025), and `Qwen3-Next-80B-A3B-Instruct` (Team, 2025). For MCQAs, evaluators assigned scores on a three-point scale reflecting answer correctness and contextual coherence. For OEQAs, a five-point scale was used to capture varying degrees of factual accuracy and completeness. Only examples meeting predefined minimum quality thresholds across evaluators were retained in the final dataset.

C CPT hyperparameters

Parameter	Value
Learning rate	2e-05 (1e-04 for gemma family)
Train batch size	2 (4 for gema family)
Seed	42
Gradient accumulation steps	2 (16 for gemma family)
Optimizer	AdamW
Weight Decay	0.01
Scheduler	Cosine
Number of epochs	3

Table 2: Hyperparameters used in CPT training

D SFT hyperparameters

Parameter	Value
Rank	16
LoRA Aplha	16
LoRA Dropout	0.05
use_dora	True
Learning rate	2e-05 (1e-04 for gemma family)
Train batch size	4
Evaluation batch size	train_batch_size * 2
Seed	42
WarmUp_ratio	0.05
Gradient accumulation steps	8
Optimizer	AdamW
Scheduler	Cosine
Number of epochs	10
Target Modules	QKVOGUD

Table 3: Hyperparameters used in SFT training

[3] https://openai.com/index/gpt-4-1/

E Preliminary Comparison of Full Fine-Tuning and PEFT

To justify our choice of parameter-efficient fine-tuning (PEFT) for SFT, we conducted preliminary experiments comparing full fine-tuning with several PEFT methods on the FrenchMedMCQA dataset (Labrak et al., 2022).

We evaluated LoRA (Hu et al., 2022), DoRA (Mao et al., 2024), and VeRA (Kopiczko et al., 2024) against full-parameter fine-tuning using identical training configurations. Results are reported in Table 4.

We observe that PEFT methods, particularly DoRA, outperform full fine-tuning while requiring significantly fewer trainable parameters. In addition, full fine-tuning exhibited higher overfitting tendencies, with faster training loss convergence but weaker generalization performance on validation data.

These results support the use of parameter-efficient methods for SFT in our main experiments, as they provide a better trade-off between performance, efficiency, and generalization.

F Evaluation Metrics

We provide here the formal definitions of the evaluation metrics used for MCQU and MCQ evaluation.

Exact Match (EM). Exact Match measures the proportion of predictions that exactly match the gold answer:

$$\text{EM} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i],$$

where N denotes the number of questions, y_i the gold answer, $\hat{y}_i$ the model prediction, and $[\cdot]$ is the indicator function.

Hamming Score. For multi-answer MCQ, we additionally report the Hamming score, which captures partial agreement between predicted and reference label sets:

$$\text{Hamming Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}.$$

This metric rewards partial correctness and is therefore better suited for evaluating multi-label predictions than Exact Match alone.

	LoRA	DoRA	VeRA	Full FT
Exact Match	0.2211	**0.2435**	0.1153	0.1121
Hamming Distance	0.4325	**0.4627**	0.3482	0.3143
Trainable Params (%)	0.583	0.602	0.0037	100

Table 4: Comparison of full fine-tuning and parameter-efficient methods on FrenchMedMCQA.

G Evaluation Benchmarks

The adapted models are evaluated against their corresponding base models using benchmark datasets drawn from the *test split* of MedInjection-FR. The evaluation suite includes both *native French* benchmarks and *translated English* benchmarks. For the translated benchmarks, English test sets were translated into French following the procedure described in section B.3.

The benchmarks cover multiple task formats, including MCQU, MCQ and OEQA. This setup enables a controlled comparison of adaptation effects across both discriminative and generative biomedical reasoning tasks. Table 5 summarizes the datasets used for evaluation and their respective sizes.

Dataset	# Items	Task
NATIVE		
	3 384	MCQ
MediQAl (Bazoge, 2025)	4 343	MCQU
	4 969	OEQA
FrenchMedMCQA (Labrak et al., 2022)	622	MCQ
mlabonne/medical-mcqa-fr[4]	150	MCQ
mlabonne/medical-cases-fr[5]	352	MCQ
FrBMedQA (Kaddari and Bouchentouf, 2022)	187	MCQU
S-Editions[6]	343	OEQA
	183	MCQ
TRANSLATED		
MedQA_4options (Jin et al., 2021)	1 273	MCQU
MedQA_5options (Jin et al., 2021)	1 273	MCQU
PubMedQA (Jin et al., 2019)	500	MCQU
MedMCQA (Pal et al., 2022)	4 183	MCQU
MMLU (Hendrycks et al., 2021)	1 080	MCQU
K-QA (Manes et al., 2024)	201	OEQA
MMLU-PRO (Wang et al., 2024b)	2 333	MCQU
MedXpertQA (Zuo et al., 2025)	2 450	MCQU

Table 5: Evaluation benchmarks used to compare adapted models with their base counterparts. All datasets correspond to the test split of MedInjection-FR.

H Prompt Templates

Overview. We use a unified instruction format across all task types, both for supervised fine-tuning and for zero-shot evaluation. When available, we rely on the native chat templates provided by instruction-tuned models; otherwise, prompts are formatted as plain-text instruction–response pairs.

Shared Structure. All prompts begin with a high-level medical instruction, optionally followed by a contextual passage. The core components are:

1. an instruction describing the task,
2. the question (and answer options when applicable),
3. an optional context section, and
4. a response header indicating where the model output should begin.

Task-Specific Constraints. The only variation across task types lies in the expected response format, which is explicitly stated in the instruction. Table 6 summarizes the templates used for each task.

Canonical Prompt Format. The following abstract template illustrates the prompt structure shared across all tasks:

System prompt (training and evaluation) for MCQ

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
Répondre uniquement avec la lettre
    correspondant à la ou les bonnes
    réponses séparées par des virgules.
    Exemple : A, C, D.
```

System prompt (training and evaluation) for MCQU

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
Répondre uniquement avec la lettre
    correspondant à la bonne réponse.
    Exemple : A.
```

Task	Instruction Constraint	Expected Output
MCQU	Respond only with the letter corresponding to the single correct answer.	Single letter (e.g., `A`)
MCQ	Respond only with the letters corresponding to all correct answers, separated by commas.	Comma-separated letters (e.g., `A, C, D`)
OEQA	Provide a free-form medical answer based on the instruction and context.	Unconstrained text

Table 6: Summary of task-specific prompt templates and output constraints.

System prompt (training and evaluation) for OEQA

```
Lire l'instruction médicale suivante et
    fournir une réponse adaptée à la
    situation décrite.
```

User prompt (task-dependent)

```
### Instruction:
[Question (+ options for MCQ tasks)]

### Contexte:
[Context, if available]

### Réponse:
```

Chat-Based Formatting. For instruction-tuned models providing an explicit chat interface, the same content is mapped to role-based messages as follows:

- **System**: high-level medical instruction (shared across tasks),
- **User**: task instruction, question, and optional context,
- **Assistant**: model-generated answer.

This formulation ensures consistent supervision and evaluation across models with different input formatting requirements.

I MCQA and OEQA Results

I.1 MCQA Greedy Decoding

Table 7 reports performance on MCQA across the studied model families (Gemma-4B, Mistral-7B, Llama-7B, and Llama-13B), three initialization types (General, Instruct, Medical), and four adaptation strategies (Base, CPT, SFT, CPT+SFT). Results are shown for both standard multiple-answer MCQs (MCQ) and single-answer MCQs (MCQU), using Exact Match (EM), Hamming score for MCQ, and aggregated EM. The reported results here are obtained using greedy decoding.

Effectiveness of Adaptation Strategy: Under greedy decoding, SFT clearly dominates all other adaptation strategies across MCQ, MCQU, and aggregated metrics. Unlike constrained decoding, where CPT+SFT often ranks first, greedy decoding exposes a much sharper separation between strategies:

$$\text{BASE} \ll \text{CPT} \ll \text{CPT+SFT} < \text{SFT}$$

Across nearly all model families and initializations, SFT yields the highest MCQ EM, MCQ Hamming, MCQU EM, and aggregated EM. This trend is particularly strong for instruction-tuned models (Mistral, Llama-7B, Llama-13B), where SFT consistently delivers large absolute gains, often by wide margins. As in the constrained decoding setting, when CPT+SFT outperforms SFT, the performance gap is generally smaller than in configurations where SFT outperforms CPT+SFT.

CPT alone remains unstable under greedy decoding. While it sometimes improves over the base model, it frequently underperforms SFT and can even degrade MCQU and aggregated scores. Importantly, CPT+SFT does not systematically improve over SFT in greedy decoding and often performs worse, indicating that the benefits of CPT are largely redundant once task supervision is introduced and decoding constraints are removed.

Overall, greedy decoding amplifies the advantages of task-aligned supervision, making SFT the best adaptation strategy when decoding is unconstrained.

Impact of Model Initialization: Model initialization plays a stronger and more consistent role under greedy decoding than under constrained decoding. Across all families and metrics, instruction-tuned models dominate. General models benefit from SFT but remain consistently below instruction-tuned counterparts. Medical models, while improving with SFT, never achieve the best

Model Type	Strategy	MCQ		MCQU	Aggregation	OEQA	
		EM	Hamming	EM	EM	*BLEU*	*METEOR*
				Gemma-4B			
GENERAL	Base	0.56	5.15	5.88	3.22	0.92	8.09
	CPT	0.03	0.68	8.53	4.28	**1.68**	**8.35**
	SFT	**1.73**	**19.48**	**19.81**	**10.77**	1.03	7.42
	CPT+SFT	1.67	15.34	19.52	10.60	0.56	6.99
INSTRUCT	Base	**6.81**	**40.75**	28.88	**17.85**	**0.76**	**10.41**
	CPT	0.07	1.72	1.37	0.72	0.46	5.89
	SFT	1.38	10.17	**31.87**	16.63	0.09	5.32
	CPT+SFT	1.22	8.61	30.66	15.94	0.21	6.93
MEDICAL	Base	1.22	11.28	11.47	6.34	0.76	**10.41**
	CPT	1.62	**18.11**	11.18	6.40	**1.04**	5.26
	SFT	1.15	11.51	**17.91**	9.53	0.56	6.98
	CPT+SFT	**1.67**	16.54	17.63	**9.65**	0.34	5.60
				Mistral-7B			
GENERAL	Base	0.15	2.90	4.04	2.09	0.66	7.32
	CPT	0.44	3.75	13.71	7.07	0.99	7.69
	SFT	**1.79**	**20.62**	**19.88**	**10.84**	**1.04**	7.59
	CPT+SFT	1.42	17.95	19.57	10.49	0.63	**8.66**
INSTRUCT	Base	3.42	26.94	21.46	12.44	1.12	7.70
	CPT	4.10	29.16	27.63	15.87	**2.34**	**10.90**
	SFT	**11.94**	**47.39**	**31.52**	**21.73**	1.65	7.08
	CPT+SFT	1.85	17.56	29.84	15.85	1.09	9.46
MEDICAL	Base	2.24	19.09	13.39	7.81	1.73	8.39
	CPT	**2.25**	18.39	12.19	7.22	**1.93**	**9.54**
	SFT	1.95	**20.68**	18.47	10.21	1.06	6.75
	CPT+SFT	1.44	16.43	**19.33**	**10.38**	1.05	8.02
				Llama-7B			
GENERAL	Base	0.17	2.69	9.46	4.82	0.49	5.91
	CPT	1.13	10.16	5.83	3.48	**1.39**	5.90
	SFT	**1.82**	**24.64**	16.08	8.95	0.51	**7.26**
	CPT+SFT	1.61	17.50	**17.11**	**9.36**	0.52	6.75
INSTRUCT	Base	0.03	0.64	0.04	0.03	0.40	2.51
	CPT	4.92	40.87	0.04	2.48	**1.72**	8.64
	SFT	**7.72**	**42.70**	**29.26**	**18.49**	1.42	6.22
	CPT+SFT	1.03	6.36	0.07	0.55	1.41	**10.07**
MEDICAL	Base	0.14	1.98	0.57	0.35	0.51	**7.06**
	CPT	**2.11**	**24.01**	12.20	7.16	**1.12**	5.69
	SFT	1.12	20.40	17.49	9.30	0.40	5.50
	CPT+SFT	1.67	15.87	**18.29**	**9.98**	0.49	6.70
				Llama-13B			
GENERAL	Base	0.29	0.84	11.77	6.03	0.19	1.98
	CPT	**2.81**	21.03	10.49	6.65	**1.85**	7.85
	SFT	2.18	17.08	18.67	10.42	0.42	5.85
	CPT+SFT	1.65	**21.56**	**19.84**	**10.74**	0.59	**8.07**
INSTRUCT	Base	0.00	4.87	0.04	0.02	0.50	3.85
	CPT	4.82	34.09	0.04	2.43	**2.09**	**10.78**
	SFT	10.92	42.98	30.13	20.53	1.52	6.52
	CPT+SFT	**12.57**	**44.02**	**31.48**	**22.02**	1.74	7.32
MEDICAL	Base	0.56	9.41	11.01	5.78	0.48	5.60
	CPT	2.02	**22.25**	11.28	6.65	**1.12**	5.69
	SFT	**2.07**	21.50	18.11	10.09	0.59	**7.37**
	CPT+SFT	1.61	17.64	**19.75**	**10.68**	0.67	6.53

Table 7: Greedy decoding results for MCQ and MCQU and BLEU/METEOR scores for OEQA across model families and adaptation strategies. **Bold** denotes the best strategy and underlining the best initialization.

greedy decoding performance, confirming that domain pretraining alone is insufficient without strong instruction alignment.

This contrasts with constrained decoding, where general and medical models occasionally remain competitive. Under greedy decoding, instruction tuning becomes a necessary condition for strong performance.

MCQA Greedy Decoding Guidelines: For greedy decoding in medical MCQA, the optimal configuration is to start from an instruction-tuned model and apply SFT only. CPT and CPT+SFT offer no consistent benefit in this setting and can be safely avoided unless constrained decoding is explicitly required.

I.2 OEQA Overlap-based Evaluation

The right part of Table 7 reports the overlap-based metrics BLEU and METEOR. Both metrics exhibit trends consistent with ROUGE-L, with improvements primarily driven by CPT. In a few isolated cases, CPT+SFT yields additional gains on METEOR, but with small differences with when compared with CPT. Regarding model initialization, BLEU and METEOR consistently favor instruction-tuned models as the strongest starting point.

J Statistical Significance

We assess whether observed differences between adaptation strategies and initialization choices are statistically significant using paired bootstrap significance testing. For each comparison, we compute the per-instance score difference (EM for MCQ/MCQU; judge-based correctness for OEQA) and report a two-sided p-value (`p_two_sided`). Statistical significance is determined by comparing this p-value against a predefined threshold α. We report results using a Bonferroni-corrected threshold to control for multiple comparisons.

For comparisons between adaptation strategies within a model family, we perform 12 pairwise tests per family, yielding a corrected threshold of $\alpha_{\text{Bonferroni}} = 0.05/12$. For comparisons between model initialization types under a fixed adaptation strategy, we perform 9 pairwise tests per family, yielding $\alpha_{\text{Bonferroni}} = 0.05/9$. The applied threshold (`alpha_Bonferoni`) and the resulting significance decision (`significant_Bonferroni`) are reported explicitly in Tables 8 and 9.

We define the mean difference as $\Delta = \text{score}(\texttt{model_a}) - \text{score}(\texttt{model_b})$ (not shown in the tables). Therefore, if the confidence interval is entirely above zero, `model_a` performs better; if it is entirely below zero, `model_b` performs better. A comparison is considered statistically significant if the corrected decision is TRUE.

J.1 Interpretation of comparison IDs

Each row in Tables 8 and 9 corresponds to a specific pairwise comparison between two models (`model_a` vs. `model_b`). The `id` field encodes the purpose of the comparison: (i) **IDs A–C** compare models *within the same model type* (GENERAL, INSTRUCT, or MEDICAL) in order to quantify the effect of adaptation strategies (CPT, SFT, CPT+SFT) relative to a fixed initialization. (ii) **IDs D–F** compare models *across model types* under a fixed adaptation strategy, in order to identify the most effective initialization point (GENERAL vs. INSTRUCT vs. MEDICAL) for downstream adaptation.

J.2 Decoding conditions

For MCQ/MCQU, Table 8 reports significance results separately for greedy and constrained decoding. For OEQA, Table 9 reports strategy-level comparisons under the evaluation setting used for the main experiments.

K Near-Miss Rates in MCQA

Model	MCQ	MCQU
Mistral	0.203	0.202
Mistral-CPT	0.212	0.203
Mistral-SFT	0.234	0.204
Mistral-CPT-SFT	0.256	0.203
Mistral-Instruct	0.173	0.202
Mistral-Instruct-CPT	0.205	0.202
Mistral-Instruct-SFT	0.174	0.206
Mistral-Instruct-CPT-SFT	0.221	0.192
BioMistral	0.181	0.205
BioMistral-CPT	0.195	0.201
BioMistral-SFT	0.220	0.211
BioMistral-CPT-SFT	0.234	0.204

Table 10: Near-miss rates for MCQ and MCQU across Mistral variants. A near-miss corresponds to cases where all gold answers are ranked within the top-k options but the generated answer does not match the gold label(s). Near-miss rates remain stable across model families and adaptation strategies, indicating that improvements in confidence and ranking do not directly translate into exact prediction.

To better characterize model behavior beyond EM accuracy in MCQA, we analyze *near-miss* predictions. Table 10 shows the near-miss rates obtained

	id	strategy	model_a	model_b	Decoding type	ci95_low	ci95_high	p_two_sided	alpha_Bonferoni	significant_Bonferroni
Gemma 4B										
GENERAL	A1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	constrained	-2.60E-02	-3.48E-03	2.00E-04	4.17E-03	**TRUE**
	A1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	greedy	-6.03E-03	2.71E-02	1.75E-01	4.17E-03	FALSE
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	constrained	1.57E-02	5.29E-02	4.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	greedy	6.03E-02	9.37E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	constrained	2.34E-02	7.52E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	greedy	4.06E-02	9.63E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	constrained	1.32E-02	4.90E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	greedy	6.17E-02	9.54E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	constrained	-4.69E-02	-8.81E-03	4.00E-03	4.17E-03	**TRUE**
	B1	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	greedy	-1.94E-01	-1.47E-01	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	constrained	-1.25E-02	1.41E-02	9.27E-01	4.17E-03	FALSE
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	greedy	-3.54E-02	-2.16E-03	2.66E-02	4.17E-03	FALSE
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	constrained	1.49E-02	4.29E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	greedy	1.26E-01	1.78E-01	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	constrained	-5.48E-03	2.13E-02	2.77E-01	4.17E-03	FALSE
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	greedy	-2.62E-02	3.58E-03	1.21E-01	4.17E-03	FALSE
MEDICAL	C1	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	constrained	-1.68E-02	1.67E-04	5.60E-02	4.17E-03	FALSE
	C1	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	greedy	-4.32E-03	9.75E-03	4.60E-01	4.17E-03	FALSE
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	constrained	2.64E-02	4.73E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	greedy	2.44E-02	3.98E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	constrained	1.72E-02	4.13E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	greedy	2.80E-02	4.25E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	constrained	1.74E-02	4.04E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	greedy	2.58E-02	4.14E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	constrained	-7.13E-03	1.15E-02	6.83E-01	5.56E-03	FALSE
	D1	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	greedy	-6.92E-02	-4.32E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	constrained	2.60E-03	1.27E-02	4.80E-03	5.56E-03	**TRUE**
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	greedy	8.83E-03	1.61E-02	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	constrained	-5.57E-03	1.57E-02	3.13E-01	5.56E-03	FALSE
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	greedy	5.63E-02	8.12E-02	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	constrained	-3.76E-02	2.01E-02	7.90E-01	5.56E-03	FALSE
	E1	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	greedy	2.58E-02	4.95E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	constrained	-2.87E-02	2.19E-02	8.35E-01	5.56E-03	FALSE
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	greedy	-5.66E-02	1.81E-03	8.56E-02	5.56E-03	FALSE
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	constrained	-2.21E-03	1.59E-02	1.83E-01	5.56E-03	FALSE
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	greedy	-9.23E-02	-3.78E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-pt-CPT-SFT	constrained	2.15E-03	2.27E-02	1.92E-02	5.56E-03	FALSE
	F1	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT-SFT	greedy	6.81E-02	-3.28E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	constrained	2.64E-03	1.80E-02	1.10E-02	5.56E-03	FALSE
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	greedy	3.47E-03	1.52E-02	1.60E-03	5.56E-03	**TRUE**
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	constrained	-1.27E-02	9.11E-03	7.58E-01	5.56E-03	FALSE
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	greedy	3.84E-02	7.95E-02	2.00E-04	5.56E-03	**TRUE**
Mistral 7B										
GENERAL	A1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	constrained	1.63E-03	2.19E-02	2.26E-02	4.17E-03	FALSE
	A1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	greedy	3.64E-02	6.00E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	constrained	3.71E-02	6.85E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	greedy	6.72E-02	1.01E-01	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	constrained	2.80E-02	5.41E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	greedy	2.82E-02	4.38E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	constrained	3.25E-02	6.88E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	greedy	7.41E-02	1.06E-01	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	constrained	1.76E-02	4.88E-02	2.00E-04	4.17E-03	**TRUE**
	B1	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	greedy	1.67E-02	5.01E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	constrained	-2.10E-02	1.47E-02	7.64E-01	4.17E-03	FALSE
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	greedy	-1.59E-02	1.76E-02	9.83E-01	4.17E-03	FALSE
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	constrained	1.29E-02	5.00E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	greedy	1.68E-02	5.21E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	constrained	2.41E-02	6.41E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	greedy	6.71E-02	1.18E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	BioMistral-7B-CPT	BioMistral-7B	constrained	-1.12E-02	1.00E-02	8.13E-01	4.17E-03	FALSE
	C1	CPT	BioMistral-7B-CPT	BioMistral-7B	greedy	-1.61E-02	2.51E-03	2.85E-01	4.17E-03	FALSE
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	constrained	2.63E-02	6.50E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	greedy	1.33E-02	4.54E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	constrained	2.23E-02	7.24E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	greedy	1.27E-02	6.02E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	constrained	1.33E-02	4.67E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	greedy	1.34E-02	4.18E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	constrained	-1.89E-02	2.68E-02	7.28E-01	5.56E-03	FALSE
	D1	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	greedy	-1.32E-01	-8.16E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	constrained	-1.41E-02	4.27E-02	4.35E-01	5.56E-03	FALSE
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	greedy	8.72E-02	1.40E-01	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	constrained	5.20E-03	2.93E-02	2.00E-03	5.56E-03	**TRUE**
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	greedy	1.57E-03	1.27E-02	1.66E-02	5.56E-03	FALSE
CPT	E1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	constrained	-5.11E-02	-3.56E-04	4.62E-02	5.56E-03	FALSE
	E1	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	greedy	-1.07E-01	-6.53E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	constrained	-9.26E-03	2.57E-02	6.18E-01	5.56E-03	FALSE
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	greedy	-1.44E-02	1.79E-02	6.45E-01	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	constrained	4.02E-03	5.94E-02	1.80E-02	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	greedy	6.80E-02	1.01E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	constrained	7.69E-03	3.25E-02	1.20E-03	5.56E-03	**TRUE**
	F1	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	greedy	-6.20E-02	-4.21E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	constrained	-9.97E-03	1.59E-02	5.02E-01	5.56E-03	FALSE
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	greedy	-1.08E-02	9.42E-03	6.69E-01	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	constrained	-3.06E-02	-1.28E-03	3.42E-02	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	greedy	3.56E-02	6.81E-02	2.00E-04	5.56E-03	**TRUE**
Llama 7B										
GENERAL	A1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	constrained	-8.99E-03	4.76E-03	6.17E-01	4.17E-03	FALSE
	A1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	greedy	-2.29E-02	-4.41E-03	3.00E-03	4.17E-03	**TRUE**
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	constrained	2.06E-02	4.76E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	greedy	4.28E-02	8.52E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	constrained	1.81E-02	4.46E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	greedy	3.37E-02	6.48E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	constrained	1.58E-02	3.86E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	greedy	3.08E-02	6.12E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	constrained	-7.23E-03	4.04E-03	8.20E-01	4.17E-03	FALSE
	B1	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	greedy	1.86E-02	3.22E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	constrained	-1.32E-02	2.43E-03	2.78E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	greedy	2.64E-03	7.96E-03	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	constrained	-7.62E-03	2.54E-04	7.10E-02	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	greedy	-2.69E-02	-1.17E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	constrained	9.28E-03	4.49E-02	2.00E-03	4.17E-03	**TRUE**
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	greedy	1.54E-01	2.19E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	meditron-7b-CPT	meditron-7b	constrained	1.19E-02	2.70E-02	2.00E-04	4.17E-03	**TRUE**
	C1	CPT	meditron-7b-CPT	meditron-7b	greedy	4.73E-02	1.05E-01	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	constrained	3.78E-02	7.33E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	greedy	6.80E-02	1.47E-01	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	constrained	2.20E-02	5.08E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	greedy	1.87E-02	4.21E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	meditron-7b-SFT	meditron-7b	constrained	3.31E-02	6.48E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	meditron-7b-SFT	meditron-7b	greedy	6.23E-02	1.36E-01	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	constrained	-3.69E-02	1.03E-02	4.19E-01	5.56E-03	FALSE
	D1	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	greedy	-1.22E-01	-6.88E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	constrained	-1.39E-02	-3.59E-03	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	greedy	-8.30E-03	1.95E-03	2.48E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	constrained	-1.92E-02	2.74E-02	9.14E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	greedy	6.34E-02	1.18E-01	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf-CPT	constrained	-2.48E-02	1.31E-03	8.28E-02	5.56E-03	FALSE
	E1	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-chat-hf-CPT	greedy	-3.69E-03	3.09E-02	1.93E-01	5.56E-03	FALSE
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	constrained	-1.69E-02	-1.09E-03	2.40E-02	5.56E-03	FALSE
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	greedy	-5.63E-02	-2.58E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	constrained	-6.40E-03	1.11E-02	6.02E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	greedy	-8.45E-02	-2.63E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	constrained	1.27E-02	3.61E-02	4.00E-04	5.56E-03	**TRUE**
	F1	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	greedy	6.09E-02	1.32E-01	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	constrained	-1.83E-02	-4.73E-03	1.00E-03	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	greedy	-1.38E-02	-1.53E-03	8.80E-03	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	constrained	-5.07E-02	-2.24E-02	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	greedy	-1.44E-01	-6.56E-02	2.00E-04	5.56E-03	**TRUE**
Llama 13B										
GENERAL	A1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	constrained	2.54E-03	2.02E-02	1.06E-02	4.17E-03	FALSE
	A1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	greedy	3.67E-03	1.98E-02	4.40E-03	4.17E-03	FALSE
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	constrained	2.37E-02	5.45E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	greedy	3.27E-02	5.90E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	constrained	1.49E-02	3.87E-02	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	greedy	2.07E-02	4.80E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	constrained	1.81E-02	4.28E-02	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	greedy	3.19E-02	5.31E-02	2.00E-04	4.17E-03	**TRUE**
INSTRUCT	B1	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	constrained	-1.18E-03	4.97E-02	6.72E-02	4.17E-03	FALSE
	B1	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	greedy	1.92E-02	3.14E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	constrained	3.87E-02	9.30E-02	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	greedy	1.82E-01	2.56E-01	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	constrained	2.54E-02	6.15E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	greedy	1.58E-01	2.31E-01	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	constrained	3.11E-02	8.21E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	greedy	1.72E-01	2.41E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	C1	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	constrained	-2.15E-02	1.77E-02	7.73E-01	4.17E-03	FALSE
	C1	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	greedy	-9.55E-03	2.34E-02	2.68E-01	4.17E-03	FALSE
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	constrained	3.11E-02	5.74E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	greedy	3.56E-02	5.96E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	constrained	2.34E-02	7.11E-02	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	greedy	1.90E-02	6.57E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	constrained	2.16E-02	4.90E-02	2.00E-04	4.17E-03	**TRUE**
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	greedy	2.84E-02	5.80E-02	2.00E-04	4.17E-03	**TRUE**
SFT	D1	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	constrained	-4.30E-02	7.32E-03	2.15E-01	5.56E-03	FALSE
	D1	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	greedy	-1.36E-01	-7.93E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	constrained	-3.65E-03	8.03E-03	4.70E-01	5.56E-03	FALSE
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	greedy	-4.03E-03	7.99E-03	5.32E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	constrained	-5.69E-03	4.60E-02	1.47E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	greedy	7.93E-02	1.41E-01	2.00E-04	5.56E-03	**TRUE**
CPT	E1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	constrained	-1.87E-02	1.02E-02	8.78E-01	5.56E-03	FALSE
	E1	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	greedy	2.31E-02	7.61E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	constrained	2.37E-03	3.99E-02	1.16E-02	5.56E-03	FALSE
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	greedy	-7.36E-03	2.25E-02	7.03E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	constrained	-5.92E-03	5.27E-02	2.03E-01	5.56E-03	FALSE
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	greedy	-5.61E-02	-2.58E-02	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	F1	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	constrained	-4.46E-02	4.48E-03	1.31E-01	5.56E-03	FALSE
	F1	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	greedy	-1.48E-01	-8.37E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	constrained	-6.90E-03	8.26E-03	9.14E-01	5.56E-03	FALSE
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	greedy	-7.39E-03	6.54E-03	7.66E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	constrained	6.57E-03	4.71E-02	1.56E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	greedy	8.45E-02	1.47E-01	2.00E-04	5.56E-03	**TRUE**

Table 8: Significance testing for MCQ/MCQU comparisons. reported separately for greedy and constrained decoding. Each row reports a paired bootstrap test between `model_a` and `model_b`. including the 95% confidence interval of the mean EM difference. the two-sided p-value. and the Bonferroni-adjusted threshold with the resulting decision. IDs A–C compare adaptation strategies within the same model type; IDs D–F compare model initializations across types.

	id	strategy	model_a	model_b	ci95_low	ci95_high	p_two_sided	alpha_Bonferoni	significant_Bonferroni
Gemma 4B									
GENERAL	**A1**	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-pt	-5.05E-02	3.69E-02	5.14E-01	4.17E-03	FALSE
	A2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt	-2.25E-01	1.65E-01	8.77E-01	4.17E-03	FALSE
	A3	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-pt-CPT	-1.74E-01	1.31E-01	8.73E-01	4.17E-03	FALSE
	A4	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-pt	-2.11E-01	1.41E-01	8.06E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	gemma-3-4b-it-CPT	gemma-3-4b-it	-4.58E-01	-2.75E-01	2.00E-04	4.17E-03	**TRUE**
	B2	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it	-4.58E-01	-1.41E-01	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	gemma-3-4b-it-CPT-SFT	gemma-3-4b-it-CPT	-4.94E-03	1.41E-01	6.44E-02	4.17E-03	FALSE
	B4	SFT	gemma-3-4b-it-SFT	gemma-3-4b-it	-3.81E-01	-2.06E-01	2.00E-04	4.17E-03	**TRUE**
MEDICAL	**C1**	CPT	medgemma-4b-pt-CPT	medgemma-4b-pt	-3.46E-01	-9.28E-02	2.00E-04	4.17E-03	**TRUE**
	C2	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt-CPT	3.71E-02	1.63E-01	2.00E-04	4.17E-03	**TRUE**
	C3	CPT+SFT	medgemma-4b-pt-CPT-SFT	medgemma-4b-pt	-2.79E-01	3.51E-02	1.20E-01	4.17E-03	FALSE
	C4	SFT	medgemma-4b-pt-SFT	medgemma-4b-pt	-1.88E-01	-1.94E-04	3.96E-02	4.17E-03	FALSE
SFT	**D1**	SFT	gemma-3-4b-pt-SFT	gemma-3-4b-it-SFT	-3.40E-02	6.17E-02	7.03E-01	5.56E-03	FALSE
	D2	SFT	gemma-3-4b-pt-SFT	medgemma-4b-pt-SFT	4.05E-02	9.53E-02	2.00E-04	5.56E-03	**TRUE**
	D3	SFT	gemma-3-4b-it-SFT	medgemma-4b-pt-SFT	2.34E-02	9.61E-02	2.00E-04	5.56E-03	**TRUE**
CPT	**E1**	CPT	gemma-3-4b-pt-CPT	gemma-3-4b-it-CPT	4.04E-03	2.43E-01	4.24E-02	5.56E-03	FALSE
	E2	CPT	gemma-3-4b-pt-CPT	medgemma-4b-pt-CPT	9.44E-02	4.14E-01	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	gemma-3-4b-it-CPT	medgemma-4b-pt-CPT	6.99E-02	1.63E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	gemma-3-4b-pt-CPT-SFT	gemma-3-4b-it-CPT-SFT	5.49E-03	7.53E-02	6.80E-03	5.56E-03	FALSE
	F2	CPT+SFT	gemma-3-4b-pt-CPT-SFT	medgemma-4b-pt-CPT-SFT	7.33E-02	1.95E-01	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	gemma-3-4b-it-CPT-SFT	medgemma-4b-pt-CPT-SFT	3.56E-02	1.55E-01	2.00E-04	5.56E-03	**TRUE**
Mistral 7B									
GENERAL	**A1**	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-v0.1	-1.74E-01	5.41E-02	6.35E-01	4.17E-03	FALSE
	A2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1	-2.09E-01	1.37E-01	8.37E-01	4.17E-03	FALSE
	A3	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-v0.1-CPT	-8.42E-02	9.06E-02	9.22E-01	4.17E-03	FALSE
	A4	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-v0.1	-2.32E-01	1.02E-01	6.17E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Mistral-7B-Instruct-v0.1-CPT	Mistral-7B-Instruct-v0.1	-4.68E-02	1.69E-01	1.48E-01	4.17E-03	FALSE
	B2	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT	-7.19E-02	-3.74E-02	2.00E-04	4.17E-03	**TRUE**
	B3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1	-1.15E-01	1.34E-01	7.76E-01	4.17E-03	FALSE
	B4	SFT	Mistral-7B-Instruct-v0.1-SFT	Mistral-7B-Instruct-v0.1	-2.64E-01	1.30E-02	1.29E-01	4.17E-03	FALSE
MEDICAL	**C1**	CPT	BioMistral-7B-CPT	BioMistral-7B	-1.74E-01	7.18E-02	6.52E-01	4.17E-03	FALSE
	C2	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B	-3.40E-02	1.03E-01	2.42E-01	4.17E-03	FALSE
	C3	CPT+SFT	BioMistral-7B-CPT-SFT	BioMistral-7B-CPT	1.04E-02	1.31E-01	9.00E-03	4.17E-03	FALSE
	C4	SFT	BioMistral-7B-SFT	BioMistral-7B	-1.41E-01	4.33E-02	4.24E-01	4.17E-03	FALSE
SFT	**D1**	SFT	Mistral-7B-v0.1-SFT	Mistral-7B-Instruct-v0.1-SFT	2.14E-02	9.83E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Mistral-7B-Instruct-v0.1-SFT	BioMistral-7B-SFT	-3.04E-02	4.71E-02	7.02E-01	5.56E-03	FALSE
	D3	SFT	Mistral-7B-v0.1-SFT	BioMistral-7B-SFT	4.48E-02	7.93E-02	2.00E-04	5.56E-03	**TRUE**
CPT	**E1**	CPT	Mistral-7B-v0.1-CPT	Mistral-7B-Instruct-v0.1-CPT	-1.51E-01	-9.28E-02	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Mistral-7B-v0.1-CPT	BioMistral-7B-CPT	-2.97E-02	1.67E-01	2.44E-01	5.56E-03	FALSE
	E3	CPT	Mistral-7B-Instruct-v0.1-CPT	BioMistral-7B-CPT	1.02E-01	3.12E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	-1.25E-01	-1.04E-02	7.40E-03	5.56E-03	FALSE
	F2	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	-2.23E-02	3.51E-02	8.31E-01	5.56E-03	FALSE
	F3	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	BioMistral-7B-CPT-SFT	4.40E-02	1.07E-01	2.00E-04	5.56E-03	**TRUE**
LLAMA-7 FAMILY									
GENERAL	**A1**	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-hf	-1.17E-01	2.53E-03	7.46E-02	4.17E-03	FALSE
	A2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf-CPT	4.46E-02	1.40E-01	2.00E-04	4.17E-03	**TRUE**
	A3	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-hf	-5.92E-02	9.36E-02	5.18E-01	4.17E-03	FALSE
	A4	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-hf	-9.58E-02	5.00E-02	8.01E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Llama-2-7b-chat-hf-CPT	Llama-2-7b-chat-hf	-1.05E-01	7.92E-02	7.39E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf	-4.91E-02	7.67E-02	4.88E-01	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT	-3.13E-02	6.42E-02	3.87E-01	4.17E-03	FALSE
	B4	SFT	Llama-2-7b-chat-hf-SFT	Llama-2-7b-chat-hf	-2.67E-01	2.10E-03	8.08E-02	4.17E-03	FALSE
MEDICAL	**C1**	CPT	meditron-7b-CPT	meditron-7b	-8.36E-03	1.95E-02	4.18E-01	4.17E-03	FALSE
	C2	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b	3.69E-03	9.40E-02	4.08E-02	4.17E-03	FALSE
	C3	CPT+SFT	meditron-7b-CPT-SFT	meditron-7b-CPT	-1.49E-02	9.99E-02	1.30E-01	4.17E-03	FALSE
	C4	SFT	meditron-7b-SFT	meditron-7b	-9.43E-02	3.43E-02	4.08E-01	4.17E-03	FALSE
SFT	**D1**	SFT	LLama-2-7b-hf-SFT	Llama-2-7b-chat-hf-SFT	-9.81E-02	-2.37E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	LLama-2-7b-hf-SFT	meditron-7b-SFT	-3.18E-02	-4.98E-03	7.80E-03	5.56E-03	FALSE
	D3	SFT	Llama-2-7b-chat-hf-SFT	meditron-7b-SFT	-4.93E-03	8.30E-02	1.49E-01	5.56E-03	FALSE
CPT	**E1**	CPT	Llama-2-7b-hf-CPT	Llama-2-7b-chat-hf-CPT	-2.96E-01	-1.31E-01	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-7b-hf-CPT	meditron-7b-CPT	-1.37E-01	-3.73E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-7b-chat-hf-CPT	meditron-7b-CPT	8.64E-02	1.72E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	LLama-2-7b-hf-CPT-SFT	Llama-2-7b-chat-hf-CPT-SFT	-2.83E-01	-5.47E-02	2.00E-04	5.56E-03	**TRUE**
	F2	CPT+SFT	LLama-2-7b-hf-CPT-SFT	meditron-7b-CPT-SFT	-7.39E-02	-3.20E-02	2.00E-04	5.56E-03	**TRUE**
	F3	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	meditron-7b-CPT-SFT	6.85E-03	2.06E-01	7.00E-03	5.56E-03	FALSE
Llama 13B									
GENERAL	**A1**	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-hf	-9.64E-02	-2.62E-02	2.00E-04	4.17E-03	**TRUE**
	A2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf	-1.25E-02	1.84E-01	1.19E-01	4.17E-03	FALSE
	A3	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-hf-CPT	6.32E-02	2.31E-01	2.00E-04	4.17E-03	**TRUE**
	A4	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-hf	-3.35E-02	9.53E-02	4.87E-01	4.17E-03	FALSE
INSTRUCT	**B1**	CPT	Llama-2-13b-chat-hf-CPT	Llama-2-13b-chat-hf	-1.03E-02	1.32E-01	1.44E-01	4.17E-03	FALSE
	B2	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf	-3.25E-01	6.12E-02	4.22E-01	4.17E-03	FALSE
	B3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT	-3.19E-01	-6.80E-02	2.00E-04	4.17E-03	**TRUE**
	B4	SFT	Llama-2-13b-chat-hf-SFT	Llama-2-13b-chat-hf	-3.78E-01	-4.44E-03	4.28E-02	4.17E-03	FALSE
MEDICAL	**C1**	CPT	MedLLaMA-13B-CPT	MedLLaMA_13B	8.59E-03	3.82E-02	9.60E-03	4.17E-03	FALSE
	C2	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA_13B	-7.65E-03	1.81E-01	1.50E-01	4.17E-03	FALSE
	C3	CPT+SFT	MedLLaMA-13B-CPT-SFT	MedLLaMA-13B-CPT	-2.00E-02	1.43E-01	1.48E-01	4.17E-03	FALSE
	C4	SFT	MedLLaMA-13B-SFT	MedLLaMA_13B	-4.70E-02	9.93E-02	3.78E-01	4.17E-03	FALSE
SFT	**D1**	SFT	Llama-2-13b-hf-SFT	Llama-2-13b-chat-hf-SFT	5.63E-03	5.57E-02	2.00E-04	5.56E-03	**TRUE**
	D2	SFT	Llama-2-13b-hf-SFT	MedLLaMA-13B-SFT	-2.49E-02	4.98E-02	6.34E-01	5.56E-03	FALSE
	D3	SFT	Llama-2-13b-chat-hf-SFT	MedLLaMA-13B-SFT	-7.22E-02	1.13E-02	6.27E-01	5.56E-03	FALSE
CPT	**E1**	CPT	Llama-2-13b-hf-CPT	Llama-2-13b-chat-hf-CPT	-4.04E-01	-2.38E-01	2.00E-04	5.56E-03	**TRUE**
	E2	CPT	Llama-2-13b-hf-CPT	MedLLaMA-13B-CPT	-9.08E-02	-6.25E-02	2.00E-04	5.56E-03	**TRUE**
	E3	CPT	Llama-2-13b-chat-hf-CPT	MedLLaMA-13B-CPT	1.63E-01	3.21E-01	2.00E-04	5.56E-03	**TRUE**
CPT+SFT	**F1**	CPT+SFT	Llama-2-13b-hf-CPT-SFT	Llama-2-13b-chat-hf-CPT-SFT	-4.38E-02	4.50E-02	8.76E-01	5.56E-03	FALSE
	F2	CPT+SFT	Llama-2-13b-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	-2.24E-02	3.90E-02	5.73E-01	5.56E-03	FALSE
	F3	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	MedLLaMA-13B-CPT-SFT	-1.39E-02	2.73E-02	5.62E-01	5.56E-03	FALSE

Table 9: Significance testing for OEQA comparisons. Each row reports a paired bootstrap test between `model_a` and `model_b`, including the 95% confidence interval of the mean difference (`ci95_low`, `ci95_high`), the two-sided p-value, and the Bonferroni-adjusted threshold (`alpha_Bonferoni`) with the resulting decision (`significant_Bonferroni`). IDs A–C compare adaptation strategies within the same model type; IDs D–F compare model initializations across types.

for MCQ and MCQU across different Mistral-based model variants and adaptation strategies.

L OEQA Evaluation: Verbosity Bias

To investigate verbosity bias in OEQA, we compute descriptive statistics of generated answer lengths across all models. Table 11 reports mean, median, and standard deviation of word and character counts over greedy OEQA outputs.

M English vs. French Benchmarks: Full Numeric Results

The main paper reports averaged results using *constrained decoding* (Figure 2). Table 12 provides the complete *numeric* EM results for both *greedy* and *constrained* decoding on the native English MCQU benchmarks (MCQU-EN) and their French translations (MCQU-FR).

M.1 Greedy decoding analysis

The greedy decoding results reported in Table 12 exhibit the same overall tendencies as those observed under constrained decoding in section 5. For the Mistral family, greedy decoding consistently yields higher performance on the French translations than on the original English benchmarks, both before and after adaptation. Conversely, Gemma and Llama models generally perform better on native English benchmarks under greedy decoding, and this advantage is preserved after French medical adaptation.

As with constrained decoding, adaptation on French medical data improves performance in both languages under greedy decoding, indicating effective cross-lingual transfer. While absolute EM scores differ between decoding strategies, greedy decoding generally producing lower scores, the relative ordering between English and French benchmarks and the direction of adaptation effects remain consistent. These results suggest that the cross-lingual patterns reported in the main paper are robust to the choice of decoding strategy.

M.2 Significance testing (English vs. French)

To assess whether the English–French performance gaps are statistically significant, we perform paired significance testing *separately for each model configuration*, i.e., for each combination of (model family/type, adaptation strategy, decoding type). For each configuration, we compute the per-item EM difference between MCQU-EN and MCQU-FR on matched translated instances, and estimate a 95% confidence interval for the mean difference together with a two-sided p-value. Because each test compares a model strictly with itself across languages and each English–French pair is independent of the others, we do not apply a Bonferroni correction. Table 13 reports the resulting confidence intervals and significance decisions.

N Effect of Translated Benchmarks on Performance and Confidence

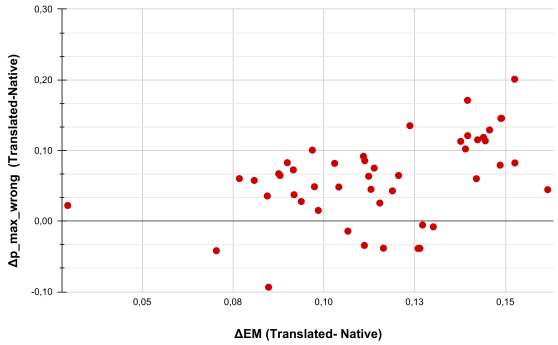


Figure 5: Relationship between accuracy gain (ΔEM) and change in confidence on incorrect predictions ($\Delta p_{\max,\text{wrong}}$) between the translated and native benchmarks. Positive values of $\Delta p_{\max,\text{wrong}}$ indicate increased confidence on incorrect predictions.

O Computational Resources and Environmental Impact

Table 14 summarizes the computational resources and environmental impact associated with each adaptation strategy, aggregated by model size. For clarity and conciseness, we do not report the consumption of each individual training run. Instead, we provide a representative summary per model size and per adaptation strategy. In total, 36 training runs were performed across all experiments. The CPT+SFT strategy is not reported as a separate entry in the table, as its computational cost and environmental impact correspond to the sum of the CPT and SFT phases. Reporting CPT and SFT independently therefore fully characterizes the overall resource usage of the combined strategy.

We report, for each configuration, the dataset size, number of epochs, batch size, GPU type, GPU memory, number of GPUs, total training time, estimated carbon emissions (in gCO_2e), and estimated monetary cost (in USD). Carbon emissions and cost estimates are derived from documented power

Model Type	Strategy	*mean_words*	*std_words*	*median_words*	*mean_chars*	*std_chars*	*median_chars*
				Gemma-4B			
GENERAL	**Base**	243,30	123,00	288,00	1 616,51	798,74	1,927,00
	CPT	107,81	122,82	37,00	720,79	806,61	245,00
	SFT	176,70	**127,77**	204,00	1 182,16	**877,20**	1 153,00
	CPT+SFT	**279,84**	87,77	**300,00**	**1 884,34**	649,64	**2 076,00**
INSTRUCT	**Base**	261,62	67,98	282,00	1 819,58	473,70	1 985,00
	CPT	**266,87**	98,66	**286,00**	1 763,62	576,08	1 878,00
	SFT	243,03	65,58	261,00	**3 513,80**	827,10	**3 326,00**
	CPT+SFT	183,16	**98,86**	207,00	3 497,45	**1 680,82**	2 632,00
MEDICAL	**Base**	208,62	122,41	205,00	1 371,82	793,35	1 431,00
	CPT	216,10	**142,15**	264,00	1 308,81	**861,75**	1 661,00
	SFT	271,22	97,47	292,00	1 796,17	705,16	**1 982,00**
	CPT+SFT	**282,47**	48,88	**283,00**	**1 825,36**	519,84	1 954,00
				Mistral-7B			
GENERAL	**Base**	212,56	58,84	224,00	1 466,40	329,83	1 502,00
	CPT	173,15	84,30	193,00	1 102,15	536,40	1,321,50
	SFT	130,36	**90,43**	138,00	884,67	**603,37**	906,00
	CPT+SFT	**226,60**	31,85	**229,00**	**1 508,78**	236,78	**1 527,00**
INSTRUCT	**Base**	134,18	74,11	125,00	876,12	476,77	812,00
	CPT	67,79	75,91	37,00	447,32	481,36	244,00
	SFT	19,59	14,19	19,00	138,75	100,30	136,00
	CPT+SFT	**168,09**	**77,91**	**199,00**	**1 112,24**	**499,35**	**1 314,00**
MEDICAL	**Base**	66,26	76,56	37,00	443,89	500,84	250,00
	CPT	99,03	**102,06**	41,00	651,69	650,72	279,00
	SFT	128,83	98,75	**159,00**	855,12	**657,31**	**950,00**
	CPT+SFT	**132,20**	91,72	129,00	**909,46**	619,60	859,00
				Llama-7B			
GENERAL	**Base**	206,13	67,03	222,00	1 358,82	392,59	**1 459,00**
	CPT	41,99	**78,06**	8,00	269,34	**490,77**	56,00
	SFT	**219,26**	35,68	220,00	**1 399,00**	266,52	1 441,00
	CPT+SFT	217,98	50,08	**224,00**	1 376,29	353,15	1 442,00
INSTRUCT	**Base**	**233,79**	66,58	**244,00**	**1 513,45**	433,69	**1 586,00**
	CPT	72,29	**87,21**	26,00	483,67	572,37	180,00
	SFT	18,24	14,33	18,00	126,03	93,60	127,00
	CPT+SFT	123,65	86,95	106,00	859,20	**581,43**	784,00
MEDICAL	**Base**	**227,34**	40,78	**230,00**	**1 498,30**	211,08	**1 522,50**
	CPT	130,92	**104,72**	128,00	847,07	**671,76**	1 008,00
	SFT	204,04	40,98	209,00	1 350,16	345,45	1 431,00
	CPT+SFT	216,96	38,84	220,00	1 416,50	301,18	1 466,00
				Llama-13B			
GENERAL	**Base**	168,15	44,63	146,00	1 144,75	260,69	1 020,00
	CPT	26,19	**53,89**	9,00	173,89	**351,27**	58,00
	SFT	218,99	36,11	219,00	**1 440,23**	300,92	**1 523,00**
	CPT+SFT	**225,12**	36,23	**228,00**	1 429,25	280,81	1 482,00
INSTRUCT	**Base**	**226,32**	60,72	**235,00**	**1 471,46**	396,55	**1 529,00**
	CPT	79,32	**80,61**	45,00	528,55	**526,68**	303,00
	SFT	17,34	10,39	18,00	121,18	74,56	129,00
	CPT+SFT	19,79	17,18	19,00	141,49	132,88	136,00
MEDICAL	**Base**	**217,49**	47,81	**221,00**	**1 429,19**	269,92	1 459,00
	CPT	65,17	**92,92**	13,00	431,03	**597,64**	92,00
	SFT	206,21	53,97	219,00	1 394,11	392,06	**1 505,00**
	CPT+SFT	215,44	37,48	220,00	1 351,46	345,39	1 402,50

Table 11: Output length statistics for OEQA generations across model families, initialization types (GENERAL/INSTRUCT/MEDICAL), and adaptation strategies (Base, CPT, SFT, CPT+SFT). We report the mean, standard deviation, and median number of words and characters per generated answer. **Bold** values highlight, within each block, the maximum value for the corresponding statistic.

Model Type	Strategy	MCQU-FR		MCQU-EN	
		Greedy	Constrained	Greedy	Constrained
		EM		EM	
Gemma-4B					
GENERAL	**Base**	**5.76**	26.63	1.60	**41.22**
	CPT	8.54	25.60	**12.77**	**40.10**
	SFT	19.92	32.82	**51.60**	**51.60**
	CPT+SFT	19.55	32.73	**51.42**	**51.42**
INSTRUCT	**Base**	29.38	29.76	**47.89**	**47.94**
	CPT	1.36	**24.28**	**1.39**	23.43
	SFT	32.38	32.46	**48.74**	**48.74**
	CPT+SFT	30.31	30.38	**39.17**	**39.17**
MEDICAL	**Base**	**11.50**	26.43	0.04	**32.47**
	CPT	**10.94**	**24.71**	7.64	23.85
	SFT	17.81	30.77	**45.03**	**45.03**
	CPT+SFT	17.41	30.50	**40.14**	**40.14**
Mistral-7B					
GENERAL	**Base**	**4.51**	**28.96**	1.34	26.15
	CPT	**13.84**	**27.15**	6.00	25.20
	SFT	**19.88**	**32.98**	7.77	27.00
	CPT+SFT	**19.47**	**32.22**	9.58	27.39
INSTRUCT	**Base**	**21.58**	**25.51**	5.96	25.10
	CPT	**28.65**	**29.69**	6.90	25.51
	SFT	**31.64**	**31.74**	7.18	26.38
	CPT+SFT	**29.94**	**30.11**	6.75	25.21
MEDICAL	**Base**	**13.52**	**26.88**	5.45	25.69
	CPT	**12.10**	**25.49**	6.05	24.32
	SFT	**18.45**	**31.64**	7.10	26.28
	CPT+SFT	**19.30**	**32.33**	7.43	26.90
Llama-7B					
GENERAL	**Base**	**9.38**	**25.48**	3.46	25.17
	CPT	6.00	25.27	**21.14**	**28.56**
	SFT	15.74	28.61	**32.86**	**32.87**
	CPT+SFT	16.97	29.77	**39.25**	**39.25**
INSTRUCT	**Base**	0.00	**24.34**	0.00	23.44
	CPT	0.00	**24.29**	0.00	23.49
	SFT	29.52	29.58	**38.73**	**38.73**
	CPT+SFT	0.00	**24.54**	0.00	23.45
MEDICAL	**Base**	0.35	**24.19**	**0.98**	23.68
	CPT	**11.97**	**25.14**	0.09	24.96
	SFT	17.38	30.27	**36.80**	**36.80**
	CPT+SFT	18.29	31.60	**36.53**	**36.53**
Llama-13B					
GENERAL	**Base**	11.20	25.51	**16.55**	**34.83**
	CPT	10.68	26.64	**29.79**	**37.21**
	SFT	17.30	30.18	**43.22**	**43.22**
	CPT+SFT	18.27	31.41	**43.62**	**43.62**
INSTRUCT	**Base**	0.00	21.68	0.00	**23.60**
	CPT	0.00	24.57	0.00	**24.85**
	SFT	30.04	30.10	**46.99**	**46.99**
	CPT+SFT	31.42	31.51	**46.57**	**46.57**
MEDICAL	**Base**	10.31	23.97	**12.48**	**24.28**
	CPT	**10.22**	23.37	10.01	**30.87**
	SFT	16.73	29.88	**37.71**	**37.71**
	CPT+SFT	18.24	31.32	**42.73**	**42.73**

Table 12: Cross-lingual comparison between native English MCQU benchmarks (MCQU-EN) and their French translations (MCQU-FR), reported as EM (%). Results are shown for both greedy and constrained decoding. For each row and decoding type, bold values indicate the higher EM between MCQU-FR and MCQU-EN.

consumption profiles and usage costs of the underlying high-performance computing infrastructure. All experiments were conducted on the Jean Zay supercomputer operated by GENCI-IDRIS[7].

O.1 Analysis

Overall, the results highlight a clear contrast between CPT and SFT in terms of computational cost and environmental impact. CPT is consistently the most resource-intensive strategy, driven by large-scale datasets, longer effective compute time, and high degrees of GPU parallelism.

In contrast, SFT incurs substantially lower emissions and monetary costs across all model sizes. This difference is not only due to the smaller dataset size, but also to the use of parameter-efficient fine-tuning: SFT is implemented with DoRA adapters rather than full weight updates, significantly reducing both memory usage and energy consumption. Despite longer wall-clock durations in some configurations, the overall compute footprint of SFT remains markedly lower than that of CPT.

As model size increases, CPT costs grow rapidly, particularly for the 13B setting, where energy consumption and carbon emissions increase sharply. SFT, while also scaling with model size, remains comparatively efficient due to its parameter-efficient design. These findings underscore the importance of adaptation strategies that balance performance gains with computational and environmental sustainability.

P Pretraining Data Contamination Study: Was NACHOS Seen During Pretraining?

Because most of the base models we evaluate (Gemma, MedGemma, Mistral, and Llama) do not disclose their full pretraining mixtures, we conducted a small contamination study to probe whether the French biomedical NACHOS corpus may have been included (or partially included) in their pretraining data. This appendix reports two complementary, lightweight detection protocols inspired by the broader literature on memorization and pretraining-data detection in large language models (Ravaut et al., 2024)

[7] http://www.idris.fr/docs/jean-zay/jean-zay/jean-zay-presentation/

Model Type	Strategy	Model	Decoding Type	ci95_low	ci95_high	p_two_sided	Significant
			Gemma-4B				
GENERAL	Base	gemma-3-4b-pt	greedy	2,57E-02	6,42E-02	2,00E-04	TRUE
	Base	gemma-3-4b-pt	constrained	-1,97E-01	-9,55E-02	2,00E-04	TRUE
	CPT	gemma-3-4b-pt-CPT	greedy	-6,57E-02	-1,71E-02	2,20E-03	TRUE
	CPT	gemma-3-4b-pt-CPT	constrained	-1,86E-01	-1,01E-01	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-CPT-SFT	greedy	-4,01E-01	-2,18E-01	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-CPT-SFT	constrained	-2,48E-01	-1,14E-01	2,00E-04	TRUE
	SFT	gemma-3-4b-pt-SFT	greedy	-3,98E-01	-2,19E-01	2,00E-04	TRUE
	SFT	gemma-3-4b-pt-SFT	constrained	-2,50E-01	-1,17E-01	2,00E-04	TRUE
INSTRUCT	Base	gemma-3-4b-it	greedy	-2,31E-01	-1,36E-01	2,00E-04	TRUE
	Base	gemma-3-4b-it	constrained	-2,27E-01	-1,33E-01	2,00E-04	TRUE
	CPT	gemma-3-4b-it-CPT	greedy	-8,82E-03	7,42E-03	9,70E-01	FALSE
	CPT	gemma-3-4b-it-CPT	constrained	-1,89E-02	4,55E-02	6,71E-01	FALSE
	CPT+SFT	gemma-3-4b-it-CPT-SFT	greedy	-1,14E-01	-6,32E-02	2,00E-04	TRUE
	CPT+SFT	gemma-3-4b-it-CPT-SFT	constrained	-1,14E-01	-6,21E-02	2,00E-04	TRUE
	SFT	gemma-3-4b-it-SFT	greedy	-2,22E-01	-9,37E-02	2,00E-04	TRUE
	SFT	gemma-3-4b-it-SFT	constrained	-2,20E-01	-9,53E-02	2,00E-04	TRUE
MEDICAL	Base	medgemma-4b-pt	greedy	6,05E-02	2,00E-01	2,00E-04	TRUE
	Base	medgemma-4b-pt	constrained	-1,05E-01	-6,63E-03	2,94E-02	TRUE
	CPT	medgemma-4b-pt-CPT	greedy	-8,53E-03	9,14E-02	1,71E-01	FALSE
	CPT	medgemma-4b-pt-CPT	constrained	-1,45E-02	4,11E-02	6,29E-01	FALSE
	CPT+SFT	medgemma-4b-pt-CPT-SFT	greedy	-2,92E-01	-1,50E-01	2,00E-04	TRUE
	CPT+SFT	medgemma-4b-pt-CPT-SFT	constrained	-1,39E-01	-4,81E-02	4,00E-04	TRUE
	SFT	medgemma-4b-pt-SFT	greedy	-3,43E-01	-1,90E-01	2,00E-04	TRUE
	SFT	medgemma-4b-pt-SFT	constrained	-1,93E-01	-8,67E-02	2,00E-04	TRUE
			Mistral-7B				
GENERAL	Base	Mistral-7B-v0.1	greedy	-1,62E-02	1,20E-01	6,98E-01	FALSE
	Base	Mistral-7B-v0.1	constrained	8,39E-02	1,70E-01	2,00E-04	TRUE
	CPT	Mistral-7B-v0.1-CPT	greedy	3,64E-02	1,48E-01	2,00E-04	TRUE
	CPT	Mistral-7B-v0.1-CPT	constrained	7,59E-02	1,43E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	greedy	5,94E-02	1,61E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-v0.1-CPT-SFT	constrained	4,90E-02	1,04E-01	2,00E-04	TRUE
	SFT	Mistral-7B-v0.1-SFT	greedy	7,77E-02	1,92E-01	2,00E-04	TRUE
	SFT	Mistral-7B-v0.1-SFT	constrained	5,74E-02	1,29E-01	2,00E-04	TRUE
INSTRUCT	Base	Mistral-7B-Instruct-v0.1	greedy	1,30E-01	1,83E-01	2,00E-04	TRUE
	Base	Mistral-7B-Instruct-v0.1	constrained	1,21E-01	1,60E-01	2,00E-04	TRUE
	CPT	Mistral-7B-Instruct-v0.1-CPT	greedy	1,74E-01	2,62E-01	2,00E-04	TRUE
	CPT	Mistral-7B-Instruct-v0.1-CPT	constrained	9,71E-02	1,73E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	greedy	1,83E-01	2,94E-01	2,00E-04	TRUE
	CPT+SFT	Mistral-7B-Instruct-v0.1-CPT-SFT	constrained	6,43E-02	1,42E-01	2,00E-04	TRUE
	SFT	Mistral-7B-Instruct-v0.1-SFT	greedy	1,92E-01	3,03E-01	2,00E-04	TRUE
	SFT	Mistral-7B-Instruct-v0.1-SFT	constrained	7,81E-02	1,56E-01	2,00E-04	TRUE
MEDICAL	Base	BioMistral-7B	greedy	4,88E-02	1,29E-01	2,00E-04	TRUE
	Base	BioMistral-7B	constrained	1,21E-01	1,63E-01	2,00E-04	TRUE
	CPT	BioMistral-7B-CPT	greedy	3,47E-02	9,30E-02	2,00E-04	TRUE
	CPT	BioMistral-7B-CPT	constrained	9,18E-02	1,20E-01	2,00E-04	TRUE
	CPT+SFT	BioMistral-7B-CPT-SFT	greedy	6,82E-02	2,06E-01	2,00E-04	TRUE
	CPT+SFT	BioMistral-7B-CPT-SFT	constrained	8,89E-02	1,88E-01	2,00E-04	TRUE
	SFT	BioMistral-7B-SFT	greedy	6,71E-02	1,94E-01	2,00E-04	TRUE
	SFT	BioMistral-7B-SFT	constrained	8,33E-02	1,68E-01	2,00E-04	TRUE
			Llama-7B				
GENERAL	Base	Llama-2-7b-hf	greedy	1,22E-02	1,24E-01	6,60E-03	TRUE
	Base	Llama-2-7b-hf	constrained	-2,49E-02	3,53E-02	8,46E-01	FALSE
	CPT	Llama-2-7b-hf-CPT	greedy	-2,23E-01	-6,14E-02	4,00E-03	TRUE
	CPT	Llama-2-7b-hf-CPT	constrained	-2,27E-02	1,19E-01	3,12E-01	FALSE
	CPT+SFT	LLama-2-7b-hf-CPT-SFT	greedy	-2,93E-01	-1,33E-01	2,00E-04	TRUE
	CPT+SFT	LLama-2-7b-hf-CPT-SFT	constrained	-1,49E-01	-3,08E-02	4,20E-03	TRUE
	SFT	LLama-2-7b-hf-SFT	greedy	-2,27E-01	-9,34E-02	4,00E-04	TRUE
	SFT	LLama-2-7b-hf-SFT	constrained	-8,38E-02	7,10E-03	8,72E-02	FALSE
INSTRUCT	Base	Llama-2-7b-chat-hf	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	Base	Llama-2-7b-chat-hf	constrained	1,84E-01	3,09E-01	2,00E-04	TRUE
	CPT	Llama-2-7b-chat-hf-CPT	greedy	0,00E+00	1,02E-04	7,23E-01	FALSE
	CPT	Llama-2-7b-chat-hf-CPT	constrained	5,27E-02	2,03E-01	2,00E-04	TRUE
	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	greedy	0,00E+00	6,80E-05	7,01E-01	FALSE
	CPT+SFT	Llama-2-7b-chat-hf-CPT-SFT	constrained	8,66E-02	1,40E-01	2,00E-04	TRUE
	SFT	Llama-2-7b-chat-hf-SFT	greedy	-1,32E-01	-3,94E-02	1,20E-03	TRUE
	SFT	Llama-2-7b-chat-hf-SFT	constrained	-1,31E-01	-3,80E-02	1,20E-03	TRUE
MEDICAL	Base	meditron-7b	greedy	-1,03E-02	-2,51E-03	1,00E-03	TRUE
	Base	meditron-7b	constrained	-1,42E-02	3,19E-02	5,00E-01	FALSE
	CPT	meditron-7b-CPT	greedy	-6,55E-03	1,08E-01	3,98E-01	FALSE
	CPT	meditron-7b-CPT	constrained	2,15E-03	6,87E-02	3,58E-02	TRUE
	CPT+SFT	meditron-7b-CPT-SFT	greedy	-2,61E-01	-7,38E-02	1,80E-03	TRUE
	CPT+SFT	meditron-7b-CPT-SFT	constrained	-1,05E-01	2,86E-02	1,86E-01	FALSE
	SFT	meditron-7b-SFT	greedy	-2,64E-01	-1,02E-01	2,00E-04	TRUE
	SFT	meditron-7b-SFT	constrained	-1,15E-01	-3,59E-03	4,00E-02	TRUE
			Llama-13B				
GENERAL	Base	Llama-2-13b-hf	greedy	-1,31E-01	5,40E-02	2,74E-01	FALSE
	Base	Llama-2-13b-hf	constrained	-1,35E-01	-4,79E-02	4,00E-04	TRUE
	CPT	Llama-2-13b-hf-CPT	greedy	-2,65E-01	-1,00E-01	2,00E-04	TRUE
	CPT	Llama-2-13b-hf-CPT	constrained	-1,09E-01	2,30E-02	1,58E-01	FALSE
	CPT+SFT	Llama-2-13b-hf-CPT-SFT	greedy	-3,31E-01	-1,52E-01	2,00E-04	TRUE
	CPT+SFT	Llama-2-13b-hf-CPT-SFT	constrained	-1,82E-01	-5,08E-02	1,60E-03	TRUE
	SFT	Llama-2-13b-hf-SFT	greedy	-3,30E-01	-1,68E-01	2,00E-04	TRUE
	SFT	Llama-2-13b-hf-SFT	constrained	-1,87E-01	-6,61E-02	2,00E-04	TRUE
INSTRUCT	Base	Llama-2-13b-chat-hf	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	Base	Llama-2-13b-chat-hf	constrained	-4,79E-02	2,61E-03	1,06E-01	FALSE
	CPT	Llama-2-13b-chat-hf-CPT	greedy	0,00E+00	0,00E+00	1,00E+00	FALSE
	CPT	Llama-2-13b-chat-hf-CPT	constrained	-2,29E-02	2,51E-02	7,65E-01	FALSE
	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	greedy	-2,09E-01	-7,94E-02	4,00E-04	TRUE
	CPT+SFT	Llama-2-13b-chat-hf-CPT-SFT	constrained	-2,06E-01	-7,95E-02	2,00E-04	TRUE
	SFT	Llama-2-13b-chat-hf-SFT	greedy	-2,21E-01	-1,15E-01	2,00E-04	TRUE
	SFT	Llama-2-13b-chat-hf-SFT	constrained	-2,20E-01	-1,16E-01	2,00E-04	TRUE
MEDICAL	Base	MedLLaMA_13B	greedy	-8,31E-02	4,03E-02	4,90E-01	FALSE
	Base	MedLLaMA_13B	constrained	-2,46E-02	1,98E-02	7,65E-01	FALSE
	CPT	MedLLaMA-13B-CPT	greedy	-3,43E-02	3,73E-02	8,95E-01	FALSE
	CPT	MedLLaMA-13B-CPT	constrained	-1,10E-01	-4,51E-02	2,00E-04	TRUE
	CPT+SFT	MedLLaMA-13B-CPT-SFT	greedy	-3,14E-01	-1,60E-01	2,00E-04	TRUE
	CPT+SFT	MedLLaMA-13B-CPT-SFT	constrained	-1,68E-01	-5,56E-02	1,40E-03	TRUE
	SFT	MedLLaMA-13B-SFT	greedy	-2,76E-01	-1,29E-01	2,00E-04	TRUE
	SFT	MedLLaMA-13B-SFT	constrained	-1,27E-01	-2,23E-02	7,20E-03	TRUE

Table 13: Paired significance testing between MCQU-EN and MCQU-FR for each model configuration (model, strategy, and decoding type). Reported values are the 95% confidence interval of the mean EM difference and the corresponding two-sided two-sided p-value; `Significant` indicates whether the difference is statistically significant. We define the difference as $(\mathrm{FR} - \mathrm{EN})$, such that positive values indicate higher performance in French.

Model Size	Strategy	Dataset size (KB)	Epochs	Batch-size	Type of GPU	Memory per GPU (GB)	Number of GPUs	Training time (hours)	Emissions (g CO2e)	Cost (USD)
4B	CPT	4 000 000	3	4	NVIDIA A100	80	24	80	49 344	1 824.62
	SFT	369	10	4	NVIDIA H100	80	3	146	11 256.6	832.48
7B	CPT	4 000 000	3	2	NVIDIA A100	80	32	40	32 896	1 216.42
	SFT	369	10	4	NVIDIA H100	80	1	190	4 883	361.12
13B	CPT	4 000 000	3	2	NVIDIA H100	80	32	100	82 240	6 082.08
	SFT	369	10	4	NVIDIA H100	80	6	122	18 812.4	1 391.27

Table 14: Summary of computational resources and environmental impact for different adaptation strategies, aggregated by model size. Reported values correspond to a representative training configuration per strategy. CPT+SFT costs are obtained by summing CPT and SFT.

P.1 Protocol 1: Prefix–Continuation Reproduction + Likelihood Heuristics

Idea. If a model has memorized (or near-memorized) training documents, conditioning on a prefix may lead it to reproduce the exact continuation, or to assign a noticeably higher likelihood to the true continuation than to a perturbed version. This is conceptually related to training-data extraction / memorization diagnostics used in prior work (Carlini et al., 2021).

Implementation. Using a sample of n=1915 NACHOS documents, we split each document into a *prefix* (first 400 characters) and a *continuation* (rest). For each sampled document, we: (i) generate up to 200 new tokens from the prefix (greedy decoding), and compute ROUGE-L between the generated continuation and the gold continuation; (ii) compute the length of the longest common prefix (LCP) between generated and gold continuations; (iii) compute the perplexity of the gold continuation conditioned on the prefix, and compare it to the perplexity of a lightly perturbed continuation (character swaps + whitespace noise), reporting the ratio $\mathrm{PPL}(\text{gold})/\mathrm{PPL}(\text{perturbed})$. We flag a case as "suspicious" if any of the following holds: ROUGE-L ≥ 0.7, LCP ≥ 200 characters, or $\mathrm{PPL}(\text{gold})/\mathrm{PPL}(\text{perturbed}) \leq 0.85$.

Results. Across models, ROUGE-L remained very low and we observed no exact continuation matches, which does *not* support verbatim memorization of long continuations under this setup. However, the fraction of items flagged as "suspicious" is extremely high (0.82–0.96), which indicates that our heuristic is likely over-sensitive (in particular, the perturbation and/or the chosen ratio threshold may dominate the flagging decision).

- Llama-2-7B: ROUGE-L $=$ 0.031, exact matches $=$ 0/1915, suspicious fraction $=$ 0.959.
- Llama-2-13B: ROUGE-L $=$ 0.020, exact matches $=$ 0/1915, suspicious fraction $=$ 0.964.
- Mistral-7B: ROUGE-L $=$ 0.014, exact matches $=$ 0/1915, suspicious fraction $=$ 0.944.
- MedGemma-4B: ROUGE-L $=$ 0.018, exact matches $=$ 0/1915, suspicious fraction $=$ 0.821.
- Gemma-3-4B: ROUGE-L $=$ 0.019, exact matches $=$ 0/1915, suspicious fraction $=$ 0.835.

Interpretation. Given the near-zero reproduction scores (ROUGE-L, exact match) but massive "suspicious" rates, this first protocol is inconclusive as a contamination detector in our setting: it does not show direct copying, and the likelihood-based heuristic is too unstable without a careful calibration procedure and stronger perturbations/-controls. This is consistent with known difficulties of turning likelihood signals into reliable membership decisions without explicit calibration (Yeom et al., 2018).

P.2 Protocol 2: DC-PDD (Divergence-based Calibration Pretraining Data Detection)

Idea. We also tested a dedicated pretraining-data detection score, DC-PDD, which estimates a per-text statistic $\beta(x)$ combining (i) the model probability of next tokens and (ii) reference token frequencies estimated from a large background corpus D' (here, French OSCAR[8]). The method is designed to be more robust than raw perplexity by incorporating a calibration term from D' (Zhang et al., 2024).

[8] https://oscar-project.org/

Implementation. For each model, we first build a tokenizer-specific unigram table $p(v; D')$ from OSCAR-FR (streaming counts, capped number of documents), then compute DC-PDD $\beta(x)$ on: (i) 1,000 NACHOS samples, and (ii) a *synthetic control* set ("non-member") of biomedical texts generated to be *unlikely* to appear in any public pretraining mixture. We report distributional statistics (median, p75/p90/p95, mean, std) for both sets and a separation diagnostic $\Delta_{\text{median}} = \text{median}(\beta_{\text{nachos}}) - \text{median}(\beta_{\text{control}})$.

Why synthetic controls? (Major limitation) Gemma-family models were released recently (June 2025), and we could not reliably curate a sufficiently large set of *web-native biomedical French texts written after the model release date* to serve as a credible "definitely-non-member" control. As a consequence, we used synthetic biomedical controls, which weakens the study: synthetic controls differ from natural corpora in style and token statistics, and thus may artificially inflate separation (or mask it), independently of membership. We therefore treat DC-PDD results as *indicative only*, not as evidence of true pretraining inclusion.

Results. DC-PDD yields consistently *lower* scores on NACHOS than on the synthetic controls (negative Δ), suggesting the models assign *slightly* more "in-distribution" likelihood structure to NACHOS than to the synthetic texts. The separation is small for Mistral/Llama and somewhat larger for Gemma/MedGemma:

- Mistral-7B: $\Delta_{\text{median}} \approx -3.23 \times 10^{-4}$.
- Llama-2-7B: $\Delta_{\text{median}} \approx -3.58 \times 10^{-4}$.
- Llama-2-13B: $\Delta_{\text{median}} \approx -2.94 \times 10^{-4}$.
- Gemma-3-4B: $\Delta_{\text{median}} \approx -8.20 \times 10^{-4}$.
- MedGemma-4B: $\Delta_{\text{median}} \approx -6.77 \times 10^{-4}$.

Interpretation. While DC-PDD produces a consistent ordering (Nachos $<$ Control), this cannot be confidently attributed to pretraining membership because our control set is synthetic and therefore not distribution-matched. In other words, the observed separation may reflect *domain/style differences* rather than exposure during pretraining. As prior work emphasizes, robust pretraining-data detection typically requires carefully constructed controls and/or calibrated baselines (e.g., Min-K% variants, calibrated likelihood tests), which we could not fully satisfy here (Zhang et al., 2024).

P.3 Summary and Takeaways

Overall, these experiments do not provide strong evidence for (or against) NACHOS being included in the undisclosed pretraining mixtures: (i) we do not observe continuation copying under our greedy prefix–continuation setup; (ii) DC-PDD shows a small but consistent separation between NACHOS and synthetic controls, but the lack of a reliable post-release, naturally occurring biomedical control corpus makes the conclusion weak. We therefore report these results for transparency, but we do not use them to support any causal claim about pretraining contamination in the main analysis.

PromptRad: Knowledge-Enhanced Multi-Label Prompt-Tuning for Low-Resource Radiology Report Labeling

Ying-Jia Lin[1], Tzu-Chin Lo[2]*, Ping-Chien Li[3], Chi-Tung Cheng[4], Chien-Hung Liao[4], Hung-Yu Kao[5]

[1]Department of Artificial Intelligence and AI Research Center, Chang Gung University,
[2]Department of Radiology, Sijhih Cathay General Hospital,
[3]Department of Medical Imaging and Intervention, Chang Gung Memorial Hospital,
[4]Department of Trauma and Emergency Surgery, Chang Gung Memorial Hospital,
[5]Department of Computer Science, National Tsing Hua University
yjlin@cgu.edu.tw, hykao@cs.nthu.edu.tw

Abstract

Automatic report labeling facilitates the identification of clinical findings from unstructured text and enables large-scale annotation for medical imaging research. Existing rule-based labelers struggle with the diverse descriptions in clinical reports, while fine-tuning pre-trained language models (PLMs) requires large amounts of labeled data that are often unavailable in clinical settings. In this paper, we propose PromptRad, a knowledge-enhanced multi-label **prompt**-tuning approach for **rad**iology report labeling under low-resource settings. PromptRad reformulates multi-label classification as masked language modeling and incorporates synonyms from the UMLS Metathesaurus into a multi-word verbalizer to enrich category representations. By fine-tuning the PLM without additional classification layers, PromptRad requires substantially less labeled data than conventional fine-tuning. Experiments on liver CT (computed tomography) reports show that PromptRad outperforms dictionary-based and fine-tuning baselines with only 32 labeled training examples, and achieves competitive performance with GPT-4 despite using a much smaller model. Further analysis demonstrates that PromptRad captures complex negation patterns more effectively than existing methods, making it a promising solution for report labeling in data-scarce clinical scenarios. Our code is available at https://github.com/ila-lab/PromptRad.

1 Introduction

Radiology reports contain valuable medical findings about patients' conditions and play a crucial role in clinical decision-making. However, these reports are usually lengthy and unstructured (see Figure 1 for an example), making it difficult to efficiently extract clinical information (Chen et al., 2018; Pons et al., 2016). Traditional approaches rely on expert knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004) and computational linguistics techniques for medical concept extraction (Aronson, 2001; Wang et al., 2017; Peng et al., 2018). However, these dictionary-based methods depend on pre-established mappings between text and medical concepts and often fail when reports use diverse or non-conventional descriptions (Irvin et al., 2019). While pre-trained language models (PLMs) such as BioBERT (Lee et al., 2019) and PubMedBERT (Gu et al., 2021) have been applied to report labeling (Wood et al., 2020; Smit et al., 2020; Li et al., 2022), fine-tuning these models requires substantial labeled data (Devlin et al., 2019; Dodge et al., 2020; Zhang et al., 2021), which is scarce in clinical settings due to the need for domain expertise during annotation.

Example Liver CT Report (de-identified):

CT study of chest and abdomen without and with IV contrast enhancement shows:

1. Rapid enlarging nodule or mass at RUL, suggestive of rapid lung metastasis.
2. Multiple new small nodules (0.2–1.0 cm) at bilateral lungs.
3. A 4cm liver mass at S5, with peripheral globular enhancement, favor hemangioma.
4. Small cysts at right lobe of liver.
5. Gallstones.

Gold Labels: Cyst ✓, Metastasis ✓, Hemangioma ✓

Figure 1: An example liver CT report from our dataset with annotated findings. Underlined terms indicate relevant mentions for each positive label.

Prompt-tuning (Gao et al., 2021; Schick and Schütze, 2021a) offers a promising alternative by transforming classification into masked language modeling, enabling PLMs to perform well with limited labeled data (Schick and Schütze, 2021c). However, existing prompt-tuning methods (Hu et al., 2022; Gao et al., 2021; Liu et al., 2023a) are designed for multi-class classification with mutually exclusive categories. In contrast, radiol-

*Work was done while the author was affiliated with Chang Gung Memorial Hospital.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 235–249
July 3-4, 2026 ©2026 Association for Computational Linguistics

ogy report labeling is inherently a multi-label task, where multiple findings may coexist in a single report. This limits the direct applicability of existing prompt-tuning approaches to report labeling.

We propose PromptRad, a knowledge-enhanced multi-label **prompt**-tuning approach for **rad**iology report labeling under low-resource settings. Here, "low-resource" refers to settings where only a limited number of expert-annotated reports are available for training. PromptRad adapts a pre-trained masked language model (Devlin et al., 2019; Gu et al., 2021) for multi-label classification through prompt-tuning. Inspired by Knowledgeable Prompt-Tuning (KPT) (Hu et al., 2022), we design a multi-word verbalizer that incorporates synonyms from the UMLS Metathesaurus (Bodenreider, 2004) as label-to-word mappings, enriching category representations for clinical contexts. We further develop automatic prompt generation based on Gao et al. (2021) to explore the space of textual templates and enhance performance.

Experiments on liver CT reports from a large medical center demonstrate that PromptRad outperforms dictionary-based methods (Aronson, 2001; Peng et al., 2018) and fine-tuning baselines, and is competitive with GPT-4 (OpenAI, 2023) even with only 32 labeled training examples. Our analysis of negation cases shows that PromptRad captures complex negation patterns more effectively than rule-based approaches, demonstrating its robustness for clinical report labeling.

In summary, our contributions are threefold:

- We formulate low-resource radiology report labeling as a multi-label prompt-tuning problem and adapt a masked language model (Devlin et al., 2019) to predict multiple clinical findings from a single report.
- We introduce a UMLS-informed multi-word verbalizer and an automatic prompt generation strategy to improve label representation and template selection in clinical contexts.
- We evaluate PromptRad on a real-world liver CT report dataset and analyze its behavior under limited supervision, including challenging negation cases.

2 Related Work

2.1 Report Labeling

Traditional approaches to report labeling rely on pre-built knowledge bases. MetaMap (Aronson, 2001) maps medical text to UMLS concepts, and NegBio (Peng et al., 2018) extends it with dependency parsing rules for negation detection. CheXpert (Irvin et al., 2019) replaces predefined concept extractors with manually curated mention lists and more sophisticated negation rules for chest X-ray (CXR) reports. However, these rule-based methods depend on predefined patterns and may fail to generalize across report types and institutions.

Machine learning approaches, including CNN-based (Majkowska et al., 2020; Shin et al., 2017) and LSTM-based methods (Dahl et al., 2021; D'Anniballe et al., 2022), have also been explored for report labeling but require large amounts of labeled data. More recently, PLM-based approaches have shown strong performance: ALARM (Wood et al., 2020) uses BioBERT (Lee et al., 2019) for MRI reports, CheXbert (Smit et al., 2020) uses BlueBERT (Peng et al., 2019) for CXR reports, and other studies have applied BERT (Devlin et al., 2019; Li et al., 2022) and domain-specific PLMs (Nowak et al., 2023) to various report labeling tasks. While effective, fine-tuning PLMs typically requires substantial labeled data (Brown et al., 2020; Dodge et al., 2020; Zhang et al., 2021), which is scarce in clinical settings. Unlike most prior radiology labelers that focus on chest X-ray reports or rely on large annotated corpora, our work focuses on low-resource multi-label labeling for liver CT reports.

2.2 Prompt-Tuning

Prompt-tuning (Schick and Schütze, 2021a; Gao et al., 2021) reformulates classification as masked language modeling, reducing reliance on newly initialized, task-specific classification layers and enabling effective few-shot learning (Schick and Schütze, 2021c). Prompt-based methods have also been explored in biomedical NLP tasks, including biomedical relation extraction (He et al., 2024) and lay summary generation (Wu et al., 2023). KPT (Hu et al., 2022) improves prompt-tuning by incorporating external knowledge into the verbalizer. Wei et al. (2022) extended prompt-tuning to multi-label text classification and proposed PTMLTC (Prompt Tuning Method for Multi-Label Text Classification), which is the closest to our work. However, PTMLTC was designed for the educational domain and relies on a single-word verbalizer, which may not capture the diverse expressions for the target categories. To the best of our knowledge, no prior work has applied prompt-

tuning to multi-label radiology report labeling under low-resource settings, especially using small local models with few-shot training.

2.3 Recent Advances in LLM-based Report Labeling

With the emergence of large language models (LLMs), recent work has explored LLM-based approaches for radiology report analysis. CheXGPT (Gu et al., 2024) uses GPT-4 as a zero-shot labeler and then trains a BERT-based model on 50k pseudo-labeled chest X-ray reports generated by GPT-4. Fytas et al. (2024) propose a multi-turn prompting strategy that combines rule-based insights with LLM predictions for chest X-ray report classification. Domain-adapted models such as Radiology-Llama2 (Liu et al., 2023b) fine-tune open-source LLMs on radiology corpora via instruction tuning, and Abdullah and Kim (2025) fine-tune a smaller LLM for report labeling on the MIMIC-CXR dataset. However, these approaches either require large volumes of (pseudo-)labeled data, domain-specific corpora for continued pretraining, or access to proprietary APIs, raising concerns regarding deployment cost and data privacy in clinical settings. In contrast, PromptRad is a prompt-tuning approach that fine-tunes a masked language model with a UMLS-informed verbalizer, requiring only 32 labeled examples and no access to external APIs.

Table 1: Number of positive findings per category in our labeled clinical report dataset. HCC: Hepatocellular Carcinoma.

Category	Training (2008–2014)	Test (2015–2017)
Cyst	275	109
HCC	214	79
Post-Treatment	205	79
Cirrhosis	168	85
Steatosis	154	93
Metastasis	101	46
Hemangioma	90	29
Total reports	773	325

3 Dataset

This study was approved by the institutional review board of the participating medical center. The dataset contains de-identified reports written in English, covering liver CT examinations from 2008 to 2017 (an example report is shown in Figure 1). The reports were annotated by two senior radiologists for the presence of seven common cancer-related findings. Each finding is labeled as positive (1) or negative (0); suspicious findings were treated as positive to avoid false negatives.

We split the data chronologically to ensure independence between training and test sets. The 2008–2014 subset contains 773 reports and serves as the candidate training pool, while the 2015–2017 subset contains 325 reports and is used as the fixed test set, as summarized in Table 1. To study low-resource scenarios, where only a small number of expert-annotated training examples are available, for each run we randomly sample K reports from the 773-report training pool using stratified sampling (Sechidis et al., 2011) to preserve the class distribution. Our goal is to evaluate how effectively different methods learn under constrained annotation scenarios, rather than to assume that all 773 reports are unavailable.

4 Method

4.1 Problem Definition

Given a radiology report r and a label space $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ with n distinct classes of findings, the report labeling task is to predict the presence of each finding $y_i \in \mathcal{Y}$ within r. Multiple findings may be present in a single report. We define the training set as $\mathcal{D}_{\text{train}} = \{(r_a, \mathcal{Y}_a)\}_{a=1}^{K}$, where K is the number of training examples and $\mathcal{Y}_a \subseteq \mathcal{Y}$ is the set of findings in report r_a.

4.2 Standard Prompt-Tuning

Prompt-tuning (Schick and Schütze, 2021a; Gao et al., 2021) fine-tunes PLMs without additional classification layers by (1) formatting the input as a cloze test with a textual prompt template containing a `[MASK]` token, and (2) defining a verbalizer that maps each class to a word in the PLM vocabulary. This allows fine-tuning with the MLM objective (Devlin et al., 2019) without introducing new parameters. We illustrate standard prompt-tuning in Figure 2 (a).

4.3 Multi-word Verbalizer of PromptRad

To identify the presence of each finding $y_i \in \mathcal{Y}$ in a report r and transform report labeling into a masked language modeling task (Devlin et al., 2019; Gao et al., 2021; Schick and Schütze, 2021b),

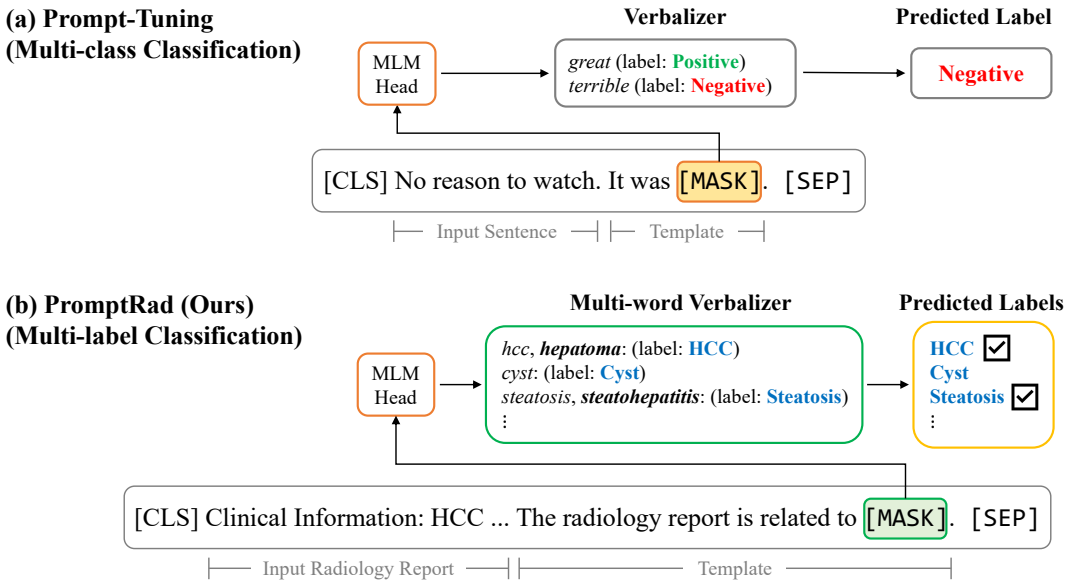


Figure 2: Differences in the approaches of (a) Prompt-tuning for multi-class classification on a general domain task (Gao et al., 2021; Schick and Schütze, 2021b). (b) PromptRad, our method for multi-label classification for radiology report labeling. Bold italic words in the verbalizer denote synonyms augmented from the UMLS Metathesaurus. MLM: masked language model (Devlin et al., 2019).

we define a verbalizer $v : \mathcal{Y} \rightarrow V$ which maps a finding y_i to a word from the vocabulary V of a pre-trained MLM $\mathcal{M}$. We first create a single-word verbalizer, where each finding is mapped to the word that most closely matches its class name (e.g., "Post-treatment" $\rightarrow$ "posttreatment"). Hence, if the model fills the word "posttreatment" in the `[MASK]` position, the report will be classified as containing the "Post-treatment" finding.

In addition to the closest match, synonyms of the class names can also represent the findings. To prevent the model from being biased towards a single mapping per category, we extend the verbalizer with synonyms from SNOMEDCT in the UMLS Metathesaurus (Bodenreider, 2004). For instance, both "hcc" and "hepatoma" serve as mappings for "Hepatocellular Carcinoma," as shown in Figure 2 (b). The complete mappings are listed in Appendix A.1. This multi-word verbalizer can be easily adapted to other report labeling tasks by replacing the mappings with synonyms for the target categories.

4.4 Training

As illustrated in Figure 3, we append a prompt template to each report r to form the input sequence r_p:

r_p = `[CLS]` r The radiology report is related to `[MASK]`. `[SEP]`

where the prompt template is manually designed. Our objective is to determine the presence of each finding by the prediction of the `[MASK]` token in r_p using a pre-trained MLM $\mathcal{M}$.

During training, for each input r_p, we identify the presence of each finding $y_i \in \mathcal{Y}$ by recovering the `[MASK]` token:

$$z_i = f_\mathcal{M}(\texttt{[MASK]}, v(y_i) \mid r_\text{p}) \quad (1)$$

where $f_\mathcal{M}(\texttt{[MASK]}, v(y_i) \mid r_\text{p})$ denotes the logit score of token $v(y_i)$ at the `[MASK]` position produced by the MLM head of $\mathcal{M}$. When multiple label-to-word mappings are available for a finding y_i, the one with the highest likelihood is chosen:

$$z_i = \max_{m=1}^{M_i}(f_\mathcal{M}(\texttt{[MASK]}, v(y_i, m)|r_\text{p})), \quad (2)$$

where M_i is the number of label-to-word mappings for y_i. For example, for "Hepatocellular Carcinoma", M_i is 2 with mappings "hcc" and "hepatoma". Practically, as we show in Figure 3, we create a matrix with the shape of $n \times M$ to facilitate the max operation in Equation 2, where M is the maximum number of mappings among the categories. We pad categories with fewer than M mappings using dummy entries and apply a binary mask so that padded positions do not affect the maximum operation. Then, we optimize $\mathcal{M}$ with the binary cross-entropy loss:

$$\hat{y}_i = \text{sigmoid}(z_i) \quad (3)$$

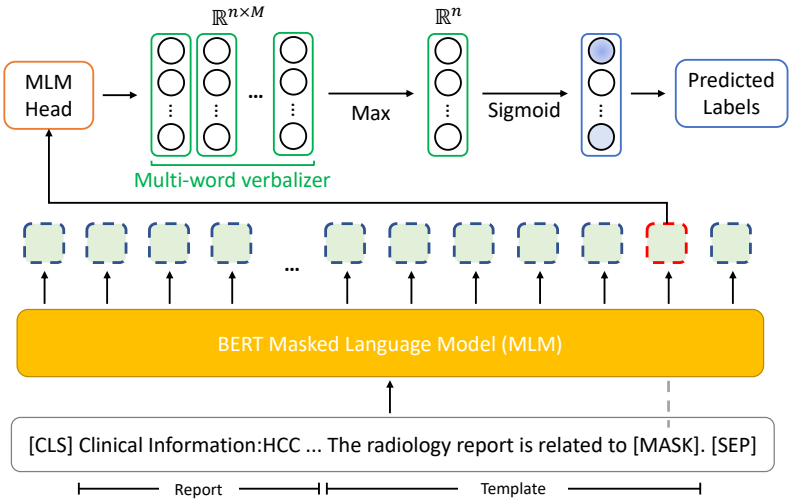


Figure 3: The workflow of the PromptRad report labeling system.

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]. \tag{4}$$

During inference, we assign a finding y_i to a radiology report r if the probability $\hat{y}_i$ exceeds a defined threshold τ for all the categories.

4.5 Automatic Prompt Generation

We argue that better performance can be achieved by training our model with different templates. Instead of relying solely on manually designed templates, we explore automatic prompt generation following Gao et al. (2021).

We take advantage of the generative capability of the pre-trained T5 (Raffel et al., 2020) to automatically create various prompt templates and find the best one for our task. Given a report r, the single-word verbalizer v, and a label y_i, we explore two formats: (1) report first: $[r]$ $[\mathbf{P}_1]$ $[v(y_i)]$ $[\mathbf{P}_2]$, and (2) answer first: $[\mathbf{P}_1]$ $[v(y_i)]$ $[\mathbf{P}_2]$ $[r]$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ are T5-generated prompts placed before and after the label word. Format 1 mirrors our manual template by placing the report before the label. In addition, since the position of the `[MASK]` token can vary in a sequence during the pre-training of BERT (Devlin et al., 2019), we also explore Format 2, where the order is reversed.

For each report with multiple labels, we duplicate the input for each label and use T5 with beam search to generate 100 candidate templates (50 per format), ranked by the joint log-likelihood over all training examples (Gao et al., 2021). We train PromptRad with each candidate and select the best-performing template (Table 6 in Appendix A.2).

5 Experiments

We evaluate PromptRad against a range of baselines on the test set. Unless otherwise specified, all training-based methods (e.g., PubMedBERT and PromptRad) use 32 labeled reports (K=32) as training data. Detailed experimental settings, baseline descriptions, and GPT-4 prompts are provided in Appendix B.

5.1 Performance Comparison

Table 2 shows that the proposed PromptRad method outperforms all the baselines for most of the categories using the manual template, except for GPT-4 (OpenAI, 2023). In addition, Table 2 also shows that PromptRad+AutoT benefits from the proposed automatic template generation method and slightly outperforms GPT-4 in terms of average scores. This result suggests that fine-tuning BERT using appropriate prompts provides a competitive and lightweight alternative to API-based large language models for low-resource report labeling.

For the category of "Hemangioma", "Label Match" can almost perfectly identify the positive findings, showing that the descriptions for "Hemangioma" can be easily identified. The reason that our approaches perform worse than "Label Match" in the "Hemangioma" category is that the number of training examples for "Hemangioma" is relatively small[1], compared to the other categories. We will

[1]During pre-processing, we randomly sampled 32 examples while keeping the distribution of the sampled data the same as that of the original training set, as we mentioned in Section 3. Thus, there are only 2-3 samples labeled as "Hemangioma" each time.

Table 2: Performance comparison in F1-score (%) using the independent test set.

	Cyst	HCC	Post-T	Cirr.	Stea.	Meta.	Hem.	Mac. F1	Mic. F1
GPT-4	86.1	91.5	79.1	96.6	98.9	73.8	95.1	88.7	88.7
GPT-4 (ICL[a])	89.7	88.8	76.3	92.5	92.0	78.5	94.9	87.5	87.4
Label Match	67.5	88.5	0.0	87.7	0.0	48.6	**98.3**	55.8	62.3
MetaMap	77.6	86.9	48.6	53.5	95.7	27.5	84.6	67.8	69.1
NegBio	77.6	86.9	**81.4**	82.7	95.7	27.5	84.6	76.6	79.2
PMB[b]	$53.7_{8.4}$	$80.4_{5.2}$	$70.4_{4.3}$	$64.5_{12.7}$	$48.3_{12.9}$	$54.9_{18.1}$	$37.7_{21.7}$	$58.6_{10.0}$	$60.9_{8.2}$
PMB[b]+MM	$68.1_{5.4}$	$83.9_{1.7}$	$64.5_{2.3}$	$68.0_{6.6}$	$84.5_{6.7}$	$47.1_{10.0}$	$70.1_{15.2}$	$69.5_{5.0}$	$70.3_{3.8}$
PMB[b]+NB	$68.1_{5.4}$	$83.9_{1.7}$	$76.8_{0.3}$	$81.0_{3.7}$	$84.5_{6.7}$	$47.1_{10.0}$	$70.1_{15.2}$	$73.1_{4.5}$	$74.1_{3.5}$
PR[b]	$78.0_{5.3}$	$89.1_{0.8}$	$76.9_{1.8}$	$86.2_{4.2}$	$95.7_{3.7}$	$71.9_{3.5}$	$88.4_{5.6}$	$83.7_{2.1}$	$84.1_{1.7}$
PR+AutoT[b]	$\mathbf{89.5}_{2.4}$	$\mathbf{90.8}_{1.0}$	$78.4_{3.9}$	$\mathbf{91.0}_{3.4}$	$\mathbf{97.3}_{0.9}$	$\mathbf{84.7}_{1.2}$	$92.4_{2.5}$	$\mathbf{89.2}_{1.0}$	$\mathbf{89.4}_{1.0}$

Abbreviations: Cirr.: Cirrhosis; Stea.: Steatosis; Meta.: Metastasis; Hem.: Hemangioma; Mac.: Macro average; Mic.: Micro average; Post-T: Post-Treatment; PMB: PubMedBERT; MM: MetaMap; NB: NegBio; PR: **P**rompt**R**ad (ours).
[a] In-Context Learning with three random samples from the training set. [b] Average from five runs; subscripts: standard deviation.

show the performance of PromptRad with more training examples in Section 5.4. We note that "Label Match" scores zero on "Post-Treatment" and "Steatosis" because these categories are expressed through synonyms (e.g., RFA, TACE, fatty liver) that do not match the category names directly.

5.2 Case Study

To confirm that the proposed method is capable of handling diverse descriptions for the target labels, we perform a case study and list the examples with the corresponding labels in Table 3. In this experiment, we compare the performance of our method with NegBio (Peng et al., 2018) as representative of the dictionary-based methods, since NegBio outperforms MetaMap (Aronson, 2001) in our experiment (Table 2). Such cases are challenging for dictionary-based systems because the target label names may appear in the text even when the corresponding findings are negated or absent.

In Table 3, we observe that most of the positive findings can be correctly identified by both approaches when the label names are explicitly mentioned in the reports. However, NegBio usually fails to negate the findings when the label names show up in reports with negative descriptions, such as "No ... nor HCC in both lobes liver" and "No imaging evidence of cirrhosis ..." in the first two examples of Table 3. Additionally, NegBio can be misled by common negation abbreviations in radiology reports, such as "R/O" (rule out) in the third report of Table 3, while learning-based methods, such as PromptRad, can learn to recognize these abbreviations from the training data.

This experiment demonstrates the potential of the proposed method in handling negation cases. To further validate the effectiveness of the proposed method for handling negation cases in radiology reports, we next perform a quantitative experiment.

Table 3: Case study comparing PromptRad+AutoT with NegBio on reports containing negated findings. Bold italic labels indicate gold annotations for each sentence.

	Segments (*Gold Label*) and Predicted Positive Labels
Report 1	A tiny hepatic cyst in S7/8> Patency of the SMV **(*Positive: Cyst*)** No obvious hypervascular nodule nor HCC in both lobes liver. **(*Negative: HCC*)** Mild liver cirrhosis and portal hypertension. **(*Positive: Cirrhosis*)** **NegBio**: Cyst, HCC, Cirrhosis **PromptRad+AutoT**: Cyst, Cirrhosis
Report 2	No imaging evidence of cirrhosis of liver. **(*Negative: Cirrhosis*)** Two small 0.6-cm and 1.4-cm densely packed lipiodol puddles in S7 without identifiable viable tumor, suggestive of good response to previous TACE without viability. **(*Positive: HCC, Post-Treatment*)** Multiple hepatic cysts are noted. **(*Positive: Cyst*)** **NegBio**: Cyst, HCC, Post-Treatment, Cirrhosis **PromptRad+AutoT**: Cyst, HCC, Post-Treatment
Report 3	Remarkable fatty liver with GB stones; **(*Positive: Steatosis*)** R/O metastasis in the anterior abdominal wall. **(*Negative: Metastasis*)** **NegBio**: Steatosis, Metastasis **PromptRad+AutoT**: Steatosis

5.3 Effectiveness of Negation Handling

To quantitatively assess the effectiveness of our proposed method in handling negation cases in radiology reports, we collect the reports from the test set that explicitly mentioned the label names, but were annotated as negative findings by our radiologists. We excluded the “Cyst", “Post-Treatment", “Steatosis", and “Hemangioma" categories, as reports mentioning these label names usually indicate positive findings of the corresponding categories.

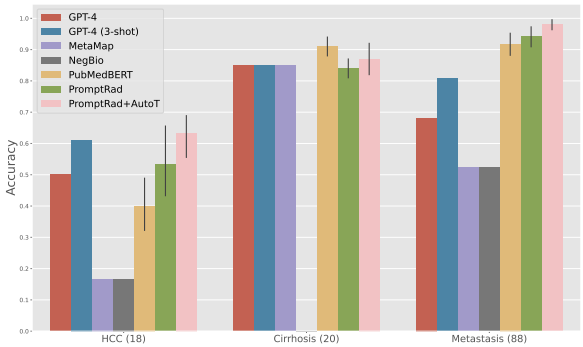


Figure 4: Accuracy on negation cases: reports that mention a finding name but are annotated as negative. Values in parentheses indicate the number of cases per category.

Figure 4 shows that PromptRad+AutoT (ours) is better at handling negation cases than GPT-4 (OpenAI, 2023), NegBio (Peng et al., 2018), and MetaMap (Aronson, 2001) for the negated reports in the three categories. Additionally, our manual-template approach (PromptRad) remains competitive compared to the other baselines. To our surprise, the performance of NegBio on “Cirrhosis" is unsatisfactory. This is due to the fact that NegBio relies heavily on natural language rules of dependency parsing and takes inputs with complete sentence structures. However, the descriptions relevant to “Cirrhosis" in our dataset are usually short (e.g., *No liver cirrhosis.*), which makes it difficult for NegBio to identify the negation cases. Besides, though our methods are slightly inferior to PubMedBERT (Gu et al., 2021) for the reports with “Cirrhosis", they are still much better at handling negation cases than PubMedBERT for the reports with “HCC" and “Metastasis". Based on the overall performance in Figure 4, we can still conclude that our method is a better choice for handling negation cases in radiology reports than the other baselines.

5.4 Performance Comparison for Different Training Sizes

In the previous sections, we studied the performance of the proposed methods for using only 32 labeled reports for training. In this section, we investigate the impact of the training size on the performance of the report labeling task. We randomly sampled different numbers of reports from the training set and evaluated the performance of PromptRad and PubMedBERT on the test set. Figure 5 shows that both methods exhibit a similar trend in performance improvement with the increase of training sizes, and both methods achieve more than 90% F1-score in macro average when the training size is larger than 128 samples. In addition, the proposed method outperforms PubMedBERT in almost all categories with training sizes smaller than 128 samples, showing that our method is effective under data scarcity scenarios.

5.5 Ablation Study for the Verbalizer

Since PromptRad and PromptRad+AutoT use the multi-word verbalizer, to confirm the effectiveness of the multi-word verbalizer, we provide a performance comparison for the two types of verbalizers in Table 4. We find that our methods using the multi-word verbalizer outperform the ones using the single-word verbalizer on average across all seven categories. Notably, the multi-word verbalizer also reduces variance across runs (e.g., Cirrhosis std. drops from 12.6 to 4.2 for PromptRad), suggesting that multiple mappings provide more stable category representations during few-shot training. Therefore, we conclude that the inclusion of multiple label-to-word mappings is beneficial for the radiology report labeling task.

We further note that, according to the complete mappings in Appendix A.1, the two verbalizers differ only for “Hepatocellular Carcinoma” and “Steatosis”. Since PromptRad is trained with shared PubMedBERT parameters and a shared MLM head, these mapping changes may indirectly affect other categories through the joint multi-label objective. Thus, differences for categories with identical mappings likely reflect shared-parameter effects. For other report labeling tasks, the multi-word verbalizer can be easily extended by adding synonyms for additional target categories when available.

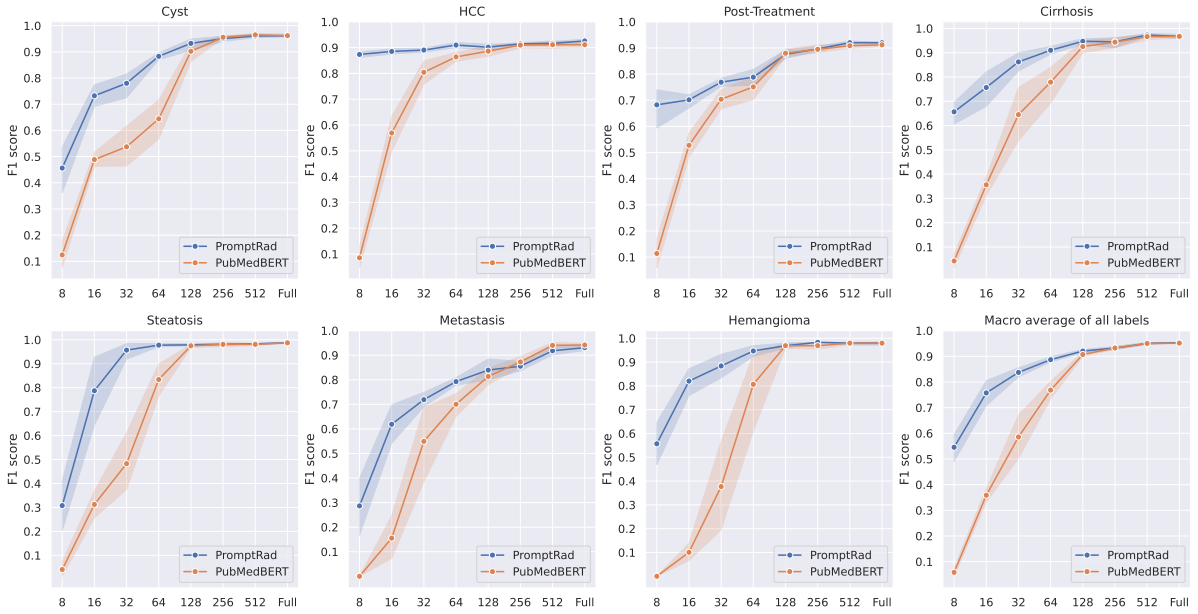


Figure 5: Comparison in F1-score for using different numbers of reports for training. We ran experiments five times for each training size, and for a training size of 8, we repeated the experiment ten times. "Full" indicates the use of the full training set.

Table 4: Performance comparison in F1-score for PromptRad and PromptRad+AutoT using the single-word and multi-word verbalizers.

	PromptRad		PromptRad+AutoT	
Labels	Single	Multi	Single	Multi
Cyst	$79.1_{4.2}$	$78.0_{5.3}$	$89.4_{3.2}$	$\mathbf{89.5}_{2.4}$
HCC	$89.4_{0.6}$	$89.1_{0.8}$	$89.9_{0.6}$	$\mathbf{90.8}_{1.0}$
Post-Treatment	$74.4_{2.4}$	$\mathbf{76.9}_{1.8}$	$\mathbf{80.2}_{3.7}$	$78.4_{3.9}$
Cirrhosis	$84.2_{12.6}$	$\mathbf{86.2}_{4.2}$	$90.3_{2.5}$	$\mathbf{91.0}_{3.4}$
Steatosis	$96.3_{2.2}$	$95.7_{3.7}$	$95.7_{2.2}$	$\mathbf{97.3}_{0.9}$
Metastasis	$74.0_{5.8}$	$71.9_{3.5}$	$83.2_{1.7}$	$\mathbf{84.7}_{1.2}$
Hemangioma	$83.6_{5.8}$	$\mathbf{88.4}_{5.6}$	$89.1_{7.8}$	$\mathbf{92.4}_{2.5}$
Macro Avg.	$83.0_{3.0}$	$\mathbf{83.7}_{2.1}$	$88.2_{1.6}$	$\mathbf{89.2}_{1.0}$
Micro Avg.	$84.1_{2.7}$	$\mathbf{84.1}_{1.7}$	$88.9_{1.4}$	$\mathbf{89.4}_{1.0}$

Scores were averaged from five runs. Subscripts: standard deviation.

6 Discussion

6.1 Why PromptRad Works

PromptRad outperforms standard BERT fine-tuning with PubMedBERT (Gu et al., 2021) across all categories when only 32 labeled reports are available (Table 2), and this advantage persists for training sizes below 128 (Figure 5). We attribute this to the alignment between prompt-tuning and masked language model pre-training. Unlike standard fine-tuning, which introduces a randomly initialized task-specific classification layer, PromptRad introduces no additional task-specific classification parameters. It reuses the pre-trained MLM head and predicts labels through verbalized clinical concepts, preserving closer alignment with the masked language modeling objective used during pre-training.

6.2 Advantages over Rule-based and API-based Alternatives

Compared to dictionary-based systems such as MetaMap (Aronson, 2001) and NegBio (Peng et al., 2018), PromptRad learns from a small number of labeled reports and better handles diverse clinical descriptions and negation patterns (Figure 4), which rule-based systems often miss due to reliance on predefined patterns.

PromptRad also achieves competitive performance with GPT-4 (OpenAI, 2023) under our low-resource setting while using a lightweight PubMedBERT backbone. Beyond efficiency, a locally deployable model avoids sending sensitive clinical reports to external APIs, which is an important consideration for real-world clinical deployment. We therefore view PromptRad not as a replacement for general-purpose large language models, but as a practical alternative when data privacy, deployment cost, and limited annotation resources are central constraints.

6.3 Comparison with Local LLMs

A natural alternative to API-based LLMs is deploying open-source LLMs locally (e.g., Llama 3

(Grattafiori et al., 2024), Qwen (Yang et al., 2025)). While this addresses privacy concerns, it requires substantial GPU memory (typically 16GB+ VRAM even for 7–8B models) that may be unavailable in smaller clinical institutions. In contrast, PromptRad uses a 110M-parameter backbone that can be deployed on consumer-grade hardware, including CPU-only inference. This makes PromptRad practical for resource-constrained clinical environments where both privacy and deployment cost are concerns.

7 Conclusion

In this paper, we introduce PromptRad, a knowledge-enhanced prompt-tuning approach for multi-label radiology report labeling under low-resource settings. Using a UMLS-based multi-word verbalizer and only 32 labeled reports, PromptRad outperforms dictionary-based and fine-tuning baselines, achieves competitive performance with GPT-4, and handles complex negation patterns more effectively than existing methods. As a lightweight, locally deployable model, PromptRad offers a practical solution for clinical report labeling where data privacy and annotation costs are central concerns, and can further support applications such as benign lesion retrieval and large-scale radiograph annotation.

Limitations

This study is based on liver CT reports from a single medical center, which may limit the generalizability of PromptRad to reports from other institutions, imaging domains, or reporting styles. In addition, our experiments focus on English reports and seven predefined liver-related findings, so further validation is needed to assess whether the method transfers well to broader clinical settings. Furthermore, the number of reports included in the negation-focused analysis is relatively small, as indicated by the counts in Figure 4. Larger annotated datasets are needed to better characterize model robustness on challenging negation cases. Finally, the construction of the multi-word verbalizer also depends on an external knowledge base such as the UMLS Metathesaurus. In settings where suitable terminology resources are unavailable, the applicability of this design may be reduced.

Acknowledgments

This work was supported by the National Science and Technology Council in Taiwan, under grants NSTC 114-2223-E-007-011 and NSTC 114-2222-E-182-001-MY2. We would like to express our deepest gratitude to the reviewers for their thoughtful comments and valuable feedback. We would also like to thank Szu-Tung Lin for his assistance in developing the data annotation platform.

References

Abdullah Abdullah and Seong Tae Kim. 2025. Automated radiology report labeling in chest x-ray pathologies: Development and evaluation of a large language model framework. *JMIR Med Inform*, 13:e68618.

AR Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings. AMIA Symposium*, page 17—21.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Matthew C. Chen, Robyn L. Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E. Chapman, David B. Larson, Curtis P. Langlotz, Timothy J. Amrhein, and Matthew P. Lungren. 2018. Deep learning to classify radiology free-text reports. *Radiology*, 286(3):845–852. PMID: 29135365.

Fredrik A Dahl, Taraka Rama, Petter Hurlen, Pål H Brekke, Haldor Husby, Tore Gundersen, Øystein Nytrø, and Lilja Øvrelid. 2021. Neural classification of norwegian radiology reports: using nlp to detect findings in ct-scans of children. *BMC Medical Informatics and Decision Making*, 21:84.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR*, abs/2002.06305.

Vincent M D'Anniballe, Fakrul Islam Tushar, Khrystyna Faryna, Songyue Han, Maciej A Mazurowski, Geoffrey D Rubin, and Joseph Y Lo. 2022. Multi-label annotation of text reports from computed tomography of the chest, abdomen, and pelvis using deep learning. *BMC Medical Informatics and Decision Making*, 22:102.

Panagiotis Fytas, Anna Breger, Ian Selby, Simon Baker, Shahab Shahipasand, and Anna Korhonen. 2024. Can rule-based insights enhance LLMs for radiology report classification? introducing the RadPrompt methodology. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 212–235, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jawook Gu, Kihyun You, Han-Cheol Cho, Jiho Kim, Eun Kyoung Hong, and Byungseok Roh. 2024. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *Preprint*, arXiv:2401.11505.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Jianping He, Fang Li, Jianfu Li, Xinyue Hu, Yi Nian, Yang Xiang, Jingqi Wang, Qiang Wei, Yiming Li, Hua Xu, and Cui Tao. 2024. Prompt Tuning in Biomedical Relation Extraction. *Journal of Healthcare Informatics Research*, 8(2):206–224.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jia Li, Yucong Lin, Pengfei Zhao, Wenjuan Liu, Linkun Cai, Jing Sun, Lei Zhao, Zhenghan Yang, Hong Song, Han Lv, and Zhenchang Wang. 2022. Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (bert) and in-domain pre-training (idpt). *BMC Medical Informatics and Decision Making*, 22:200.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, and 2 others. 2023b. Radiology-llama2: Best-in-class large language model for radiology. *Preprint*, arXiv:2309.06419.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, and 1 others. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431.

S Nowak, David Biesner, YC Layer, M Theis, Helen Schneider, W Block, Benjamin Wulff, UI Attenberger, Rafet Sifa, and AM Sprinkart. 2023. Transformer-based structuring of free-text radiology report databases. *European Radiology*, 33(6):4228–4236.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings*, 2017:188—196.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. 2016. Natural language processing in radiology: A systematic review. *Radiology*, 279(2):329–343. PMID: 27089187.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021c. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bonggun Shin, Falgun H. Chokshi, Timothy Lee, and Jinho D. Choi. 2017. Classification of radiology reports using neural attention models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4363–4370.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, Los Alamitos, CA, USA. IEEE Computer Society.

Liting Wei, Yun Li, Yi Zhu, Bin Li, and Lejun Zhang. 2022. Prompt tuning for multi-label text classification: How to link exercises to knowledge concepts? *Applied Sciences*, 12(20).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, and Thomas C. Booth. 2020. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 811–826. PMLR.

Yu-Hsuan Wu, Ying-Jia Lin, and Hung-Yu Kao. 2023. IKM_Lab at BioLaySumm task 1: Longformer-based prompt tuning for biomedical lay summary generation. In *Proceedings of the 22nd Workshop*

on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 602–610, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

A Appendix

A.1 Verbalizer Mappings

As shown in Table 5, the differences between the single-word and multi-word verbalizers fall into the categories of "Hepatocellular Carcinoma" and "Steatosis", as there are no additional synonyms for the remaining categories.

A.2 Automatically Generated Templates

We list the automatically generated templates with the highest scores in Table 6. We observe that the T5 model (Phan et al., 2021) tends to generate short but concise prompts, such as "Hepatic `[MASK]`" or "Liver `[MASK]`", while the manual template (also listed in Table 6) is more descriptive.

B Implementation Details

We implement our model using the HuggingFace Transformers (Wolf et al., 2020) (version: 4.12.2) and the PyTorch (Paszke et al., 2019) (version: 1.10.0) libraries. We used NVIDIA GeForce GTX 1070 and RTX 2080 Ti GPUs for training our model and finding templates with automatic prompt generation under the CUDA version of 10.1.

To ensure fair comparisons, all experiments for PromptRad and PubMedBERT use the model checkpoint of the pre-trained PubMedBERT-base[2] (Gu et al., 2021) for initialization, and we use SciFive-base[3] (Phan et al., 2021) for automatic prompt generation. For optimizing the models, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer to train the models. All experiments are reported with 32 labeled reports as training data, unless otherwise specified. To simulate the real-world scenario of only limited labeled data available, we do not use a development set for hyperparameter tuning. Instead, we choose the best configuration according to the training loss. We use grid search with the following ranges:

- Learning rate: $[2e-5, 3e-5, 5e-5]$
- Batch size: $[2, 4, 8]$
- Threshold τ: $[0.2, 0.3, 0.4, 0.5]$
- Ratio of training steps for learning rate warmup: $[0.0, 0.1]$

B.1 Baseline Methods

This section describes the baseline methods we used for comparison in our experiments.

B.1.1 GPT-4

We used the OpenAI API[4] to test GPT-4 on report labeling, with the model name[5] of "gpt-4". Our input instructions can be found in Table 7, where they are displayed as both the system message and the user message. For each liver CT report, we first replaced **[REPORT]** in the user message with the actual report text. The modified user message was then processed by GPT-4 to produce the corresponding output assistant message.

B.1.2 GPT-4 (ICL)

The use of in-context learning (ICL) (Brown et al., 2020; OpenAI, 2023) helps large language models achieve better performance by providing a few labeled examples in an input prompt. In this study, we also compare our approach with GPT-4 (ICL). We provide three random examples (3-shot) from the training set in the user message (Table 7) to the GPT-4 model to see if the performance can be improved through in-context learning.

B.1.3 Label Match

We create a straightforward rule-based baseline that identifies positive findings when a report contains the corresponding label names. Our matching rules also account for plural forms of label names. For instance, an occurrence of "metastases" in a report is also recognized as a positive finding for "metastasis". This simple baseline can also be used to observe the difficulties of each category.

[2] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract

[3] https://huggingface.co/razent/SciFive-base-Pubmed_PMC

[4] https://openai.com/blog/openai-api

[5] https://platform.openai.com/docs/models

Table 5: The label-to-word mappings (verbalizer) we used for PromptRad and PromptRad+AutoT.

Category	Single-Word Verbalizer	Multi-Word Verbalizer
Cyst	cyst	cyst
Hepatocellular Carcinoma	hcc	hcc, **hepatoma**
Cirrhosis	cirrhosis	cirrhosis
Post-Treatment	posttreatment	posttreatment
Steatosis	steatosis	steatosis, **steatohepatitis**
Metastasis	metastasis	metastasis
Hemangioma	hemangioma	hemangioma

Table 6: The top-5 templates from automatic template generation along with the manual template for comparison. "Manual" was used in PromptRad, while "AutoT[1]" was used in PromptRad+AutoT.

Template	Format
Manual	[*Report*] The liver radiology report is related to `[MASK]`.
AutoT[1]	Hepatic `[MASK]`: [*Report*]
AutoT[2]	Liver `[MASK]`: [*Report*]
AutoT[3]	[*Report*] Hepatic `[MASK]`.
AutoT[4]	Abdominal `[MASK]`. [*Report*]
AutoT[5]	Liver `[MASK]` in [*Report*]

B.1.4 MetaMap

MetaMap (Aronson, 2001) is a widely used dictionary-based tool for extracting concepts from biomedical text by mapping text to the UMLS Metathesaurus (Bodenreider, 2004). Due to its vast repository of biomedical knowledge and its integration of both hierarchical and non-hierarchical relationships using semantic types and relations, MetaMap stands as a fundamental and effective benchmark for report labeling tasks (Wang et al., 2017). We use the 2020 version of MetaMap with pymetamap[6], an open source MetaMap wrapper, to extract the concepts for each report in our test set.

For each report, we obtain a positive finding if MetaMap returns one of the corresponding UMLS CUIs (Concept Unique Identifiers) of each category. The mappings between CUIs and categories that we used for MetaMap to retrieve the positive findings in our task are provided in Table 8.

B.1.5 NegBio

NegBio (Peng et al., 2018) is a dictionary-based tool for negation and uncertainty detection in clinical text. This approach should work in conjunction with either CheXpert (Irvin et al., 2019) or MetaMap (Aronson, 2001). We use the official implementation of NegBio[7] to extract the concepts for each report in our test set using the same CUIs in Table 8.

B.1.6 PubMedBERT

We fine-tune the pre-trained PubMedBERT model (Gu et al., 2021) for multi-label classification using the HuggingFace API (Wolf et al., 2020). This baseline serves as a representative of the standard fine-tuning approach for PLMs. Similar to the proposed method, we set a threshold τ to determine the positive findings for each label. We perform hyperparameter tuning for PubMedBERT with the same hyperparameter ranges as the proposed method, as described in Section B.

B.1.7 PubMedBERT+MM

We create an ensemble of PubMedBERT (Gu et al., 2021) and MetaMap (MM) (Aronson, 2001) by combining the predicted positive findings from PubMedBERT and MetaMap, denoted "PubMedBERT+MM". This baseline serves as a hybrid model that combines the strengths of both expert knowledge from MetaMap and the machine learning ability of PubMedBERT.

[6] https://github.com/AnthonyMRios/pymetamap

[7] https://github.com/ncbi-nlp/NegBio

B.1.8 PubMedBERT+NB

In addition to PubMedBERT+MM, we also create "PubMedBERT+NB", an ensemble of PubMedBERT (Gu et al., 2021) and NegBio (NB) (Peng et al., 2018). As with PubMedBERT+MM, PubMedBERT+NB combines the predicted positive findings from PubMedBERT and NegBio to take advantage of both methods.

B.2 GPT-4 Prompt Design

Table 7 shows the input message we used to query GPT-4 via the OpenAI API for report labeling. The system message assigns the role of a radiologist, and the user message provides the classification task along with the report text. For GPT-4 (ICL), we additionally include three randomly sampled labeled examples from the training set in the user message.

B.3 UMLS Concept Identifiers

Table 8 lists the UMLS Concept Unique Identifiers (CUIs) used with MetaMap and NegBio to identify positive findings for each category. Since "Post-Treatment" encompasses multiple treatment procedures, we consider a report positive for this category if any of the corresponding CUIs is detected.

B.4 Evaluation Metric

We use the F1-score as the evaluation metric for the seven categories in our experiments. We also include the micro-averaged F1-score and macro-averaged F1-score for the overall performance of each model. All the experiments use the same test set mentioned in Section 3 for evaluations. For the training-based methods, including PubMedBERT (Gu et al., 2021) and our approaches, we report the average scores and standard deviations of 5 runs with different random seeds. For the experiment with only eight training examples, we report the results of 10 runs with different random seeds. This is because the performance of the training-based methods can be more unstable when the training size is smaller, and we want to provide a more comprehensive evaluation for this experiment.

Table 7: Input message and an example output of GPT-4 using the OpenAI API.

Role	**Content**
System	You are a professional radiologist who knows computed tomography (CT) very much. You can classify the CT report for the liver features or symptoms.
User	Now you are going to perform a multi-label classification for a text report of liver computed tomography (CT). Given the potential categorized features {'cyst': 0, 'HCC': 1, 'post-treatment': 2, 'cirrhosis': 3, 'steatosis': 4, 'metastasis': 5, 'hemangioma': 6}, please read the following liver computed tomography report of a patient. If the report is positive with a feature, please return the corresponding value. Liver computed tomography: **[REPORT]**
Assistant Response (Example)	1. HCC (Hepatocellular carcinoma) 3. Cirrhosis

Table 8: Concept Unique Identifiers (CUIs) used with MetaMap (Aronson, 2001) and NegBio (Peng et al., 2018) for our task. RFA: Radiofrequency Ablation; TACE: Transarterial Chemoembolization.

Category	**CUI Term (CUI)**
Cyst	Liver cyst (C0267834)
HCC	Liver carcinoma (C2239176)
Post-Treatment	RFA (C0850292)
	TACE (C3539919)
	Embolization, Therapeutic (C0013931)
	Lobectomy (C0023928)
Cirrhosis	Liver Cirrhosis (C0023890)
Steatosis	Fatty Liver (C0015695)
Metastasis	Metastasis (C4255448)
Hemangioma	Hemangioma (C0018916)

Diagnosing Lower Extremity Arteriovenous Diseases Using Agentic LLMs

Zicen Liao[1] **Yunhao Sun**[2] **Matthew Purver**[1,3]
[1]Queen Mary University of London
[2]The First Affiliated Hospital of Army Medical University, PLA
[3]Jožef Stefan Institute
z.liao@qmul.ac.uk sunyh25@alumni.sysu.edu.cn m.purver@qmul.ac.uk

Abstract

Large language models (LLMs) achieve strong results on medical benchmarks, but real outpatient diagnosis is iterative and requires stable, guideline-consistent reasoning rather than one-shot predictions. We study this problem in lower-extremity arteriovenous (LEA) diseases and introduce LEA-Dialog, a multi-turn diagnostic dialogue dataset covering six common diseases, with stage annotations for each turn and guideline-grounded probability trends. Building on this dataset, we further propose a process-aligned agentic framework for real-world patient consultations. In this framework, the patient interacts with a Doctor agent, while a Teacher agent audits guideline adherence and reasoning quality, provides corrective feedback to refine the Doctor agent's behaviour, and ultimately produces a structured diagnostic report. An Evaluator agent then derives both turn-level and final rewards from the diagnostic conversations between the patient and the Doctor agent for PPO-based policy optimisation, while carefully selected high-quality diagnostic conversations are further used for supervised fine-tuning. For training, the patient is also simulated by a Patient agent that emulates real patient responses. In addition to final diagnostic accuracy, we evaluate unsupported claims, trend correctness, diagnostic-tree similarity, and repeated-run stability with a new metric, Rate-n. Experiments on both online and offline LLMs show that process-aligned multi-turn diagnosis reduces reasoning drift, improves stability, and yields the largest gains for smaller offline models.

1 Introduction

Large language models (LLMs) have achieved strong performance on a wide range of medical benchmarks, including medical question answering, report generation, and simulated diagnosis. However, high benchmark accuracy alone does not make a system clinically usable. In outpatient practice, diagnosis is rarely a one-shot prediction: physicians iteratively summarise the current patient state, refine the differential diagnosis, ask targeted follow-up questions, and update their beliefs as new evidence emerges (Vessey and Galletta, 1991). In this setting, clinicians care not only about final correctness but also about stable and guideline-consistent reasoning (Yang et al., 2024; Rosenbacke et al., 2024).

This gap is especially important for multi-turn diagnostic dialogue. A model may produce the correct final diagnosis while still showing undesirable behaviour during the interaction, such as drifting between incompatible hypotheses, asking poorly targeted questions, or generating unsupported intermediate claims. Such process-level failures reduce trust and make the model's reasoning difficult to inspect in clinical workflows. Nevertheless, most existing evaluations of medical LLMs still focus primarily on final diagnostic accuracy, with limited attention to turn-level reasoning quality, diagnostic trajectory consistency, or repeated-run stability (Liu et al., 2024; Schmidgall et al., 2024; Fan et al., 2025; Wang et al., 2024; Lyu et al., 2024; Xu et al., 2024). In our setting, even strong online models can show noticeable variation across repeated runs on the same case, suggesting that stability should be treated as a first-class target in medical dialogue evaluation.

To study this problem, we focus on lower-extremity vascular diseases, where diagnosis often depends on multi-round clarification of symptoms, progressive narrowing of the differential diagnosis, and dynamic updates of disease likelihoods across consultation turns. Guided by clinical reasoning practice, we formulate outpatient diagnosis as a process-aligned hypothetico-deductive loop consisting of four stages at each turn: Summary, Inference, Question Planning, and Diagnosis. Based on this formulation, we introduce LEA-Dialog, a multi-turn diagnostic dialogue dataset covering

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 250–267
July 3-4, 2026 ©2026 Association for Computational Linguistics

six common lower-extremity arterial and venous diseases. Each dialogue is annotated with turn-wise reasoning and guideline-grounded probability trends, enabling evaluation of both final outcomes and intermediate reasoning behaviour.

Building on this dataset, we propose a process-aligned agentic framework for multi-turn diagnosis. A Doctor agent conducts the consultation, a Patient agent simulates the case trajectory, a Teacher agent audits record consistency, guideline adherence, and inquiry quality to construct a structured report, and an Evaluator agent converts these signals into turn-level and final rewards for optimisation. We also introduce Rate-n, a repeated-run stability metric that measures whether a model follows the same diagnostic trajectory across the first n turns. Specifically, a trajectory is considered consistent if its trend direction matches the majority trend across runs at each turn. Experiments on both online and offline LLMs show that process alignment reduces reasoning drift, improves diagnostic stability, and yields larger gains for smaller offline models.

Our contributions are as follows:

1. We introduce LEA-Dialog, a multi-turn diagnostic dialogue dataset covering six common diseases, annotated with reasoning processes for each turn and a carefully curated diagnostic handbook.

2. We propose a process-aligned agentic diagnostic framework with Doctor, Patient, Teacher, and Evaluator roles, together with a turn-level optimisation strategy for stabilising multi-turn reasoning.

3. We show that, under our framework, o3 achieves 4.6% improvement in average diagnostic accuracy and 66.7% reduction in misdiagnosis rate.

4. We further show that Qwen3-14B-T gains 68.3% in average diagnostic accuracy under the proposed framework, surpassing the original o3 model by 0.9%.

5. In addition to diagnostic accuracy, we demonstrate substantial gains in reasoning stability: after applying our framework, Qwen3-14B-T exceeds o3 in output stability by 5% from the second turn onward and by 25% from the third turn onward.

2 Related Work

2.1 Medical LLMs and diagnostic dialogue

Large language models have shown strong performance on medical question answering and benchmark-style evaluation, including systems such as AMIE (Tu et al., 2025), Med-PaLM2 (Singhal et al., 2025), and recent medical benchmarks in English and Chinese (Liu et al., 2024; Schmidgall et al., 2024; Fan et al., 2025; Wang et al., 2024; Lyu et al., 2024; Xu et al., 2024). More recent work has also begun to move from static question answering toward interactive diagnosis and simulated consultation, such as multi-agent clinical environments and conversational diagnostic systems (Schmidgall et al., 2024; Fan et al., 2025; Tu et al., 2025; Liu et al., 2025). However, most prior work still evaluates models primarily by final-answer quality, with limited focus on intermediate reasoning consistency.

2.2 Process-aligned and multi-step medical reasoning

A related line of work attempts to align model behaviour with the structure of clinical reasoning. In particular, Reasoning Like a Doctor (Xu et al., 2024) formulates medical dialogue generation through diagnostic reasoning process alignment, and MSDiagnosis introduces EMR-based multi-step diagnosis annotations (Hou et al., 2024). These studies support the view that medical dialogue should be modelled as a staged reasoning process rather than a single response generation problem. Our work follows this direction, but focuses more explicitly on the iterative nature of outpatient diagnosis and on modelling the diagnostic process itself rather than only the final prediction. Specifically, we treat diagnosis as a multi-round reasoning loop in which the model summarises accumulated evidence, updates its differential diagnoses, determines what information is still missing, and asks targeted follow-up questions to reduce uncertainty before making a final judgment. We further model turn-level diagnostic trajectory evolution, requiring disease hypotheses to change in a clinically meaningful manner across rounds. To support this, we introduce guideline-grounded supervision that specifies how diagnostic probabilities should be updated as new symptoms, history, and examination findings emerge, enabling process-level training and evaluation beyond endpoint accuracy. The comparison of evaluation dimensions between our model and other approaches is shown in Table 1.

2.3 Trustworthiness, guideline adherence, and stability

Accuracy alone is not sufficient for evaluating medical AI systems. Prior work has also highlighted the

Model	Focus	Stability
AMIE	Diagnostic dialogue	Not central
Med-PaLM2	QA and safety	Not central
ChatDoctor	Response quality	Not central
DocCHA	Online diagnosis	Not central
Emulation	Process alignment	Not central
Ours	Turn-level trajectory	Rate-n

Table 1: Evaluation focus of representative studies: AMIE (Tu et al., 2025), Med-PaLM2 (Singhal et al., 2025), ChatDoctor (Li et al., 2023), DocCHA (Liu et al., 2025), and Emulation (Xu et al., 2024). Stability refers to repeated-run diagnostic trajectory stability; "Not central" indicates that it is not a primary evaluation target.

importance of fitting clinical workflows (Xu et al., 2024), providing interpretable reasoning, and supporting clinician trust. In outpatient settings, diagnosis is inherently iterative: clinicians summarise the case, update differential diagnoses, ask targeted follow-up questions, and gradually move toward a decision. Yet the stability of this process across repeated runs remains rarely examined in medical AI, as shown in Table 1. Our work addresses this gap by studying multi-turn LLM diagnosis from a process perspective, introducing a process-aligned framework, a dataset with turn-level trend supervision, and a repeated-run stability metric, Rate-n.

3 LEA-Dialog Dataset

3.1 Task Scope

LEA-Dialog focuses on six common lower-extremity arteriovenous (LEA) diseases: deep vein thrombosis (DVT), post-thrombotic syndrome (PTS), arteriosclerosis obliterans (ASO), arterial thrombosis (AT), varicose veins (VV), and thromboangiitis obliterans (Buerger disease). We choose this disease set because diagnosis in this domain is highly iterative, requiring repeated clarification of symptoms and progressive updating of disease likelihoods across consultation rounds.

3.2 Data Composition

The dataset contains three components:

1. Expert-written consultation records. We collect 120 consultation records manually compiled and refined by experts based on real outpatient visits.

2. Synthetic consultation records. To improve coverage and diagnostic difficulty, we further construct 480 synthetic consultation records using data augmentation based on expert-crafted seed cases.

3. Clinical guideline file. In addition to the dialogue data, we include a companion guideline document for lower-extremity vascular medicine, which specifies diagnostic criteria and structured probability-scoring rules for supervising disease-trend evolution across turns.

Overall, LEA-Dialog combines expert-authored high-quality cases with larger-scale augmented data, while keeping all dialogues grounded in a shared guideline source. The detailed case collection procedure is provided in Appendix A, and the handbook development plan and reference guidelines are presented in Appendix B.

3.3 Annotation Scheme

Each dialogue contains up to 15 turns, with most expert cases reaching a final diagnosis within 5-7 rounds. For each turn, the doctor-side reasoning is annotated with four process-aligned stages:

Summary: a structured summary of the current patient state;

Inference: revision of the differential diagnosis based on newly acquired evidence;

Question Planning: targeted follow-up questions for reducing diagnostic uncertainty;

Diagnosis: final diagnostic output, produced only when sufficient evidence has been collected.

In the final round, each case additionally includes a trend report, which records how the estimated probability of each candidate disease changes across turns and why. Unlike free-form confidence scores, these trends are grounded in the companion guideline file, making them suitable for both evaluation and reward design in later training.

Compared with prior multi-turn medical datasets, LEA-Dialog provides stage annotations for each turn, guideline-grounded probability trends, and an expert-developed diagnostic handbook, enabling evaluation of diagnostic trajectory quality and repeated-run stability. Specific examples from the dataset are provided in Appendix C.3, Appendix C.5, and Appendix C.6.

3.4 Data Construction

We construct LEA-Dialog in three stages. First, three researchers with vascular-surgery backgrounds expand de-identified and standardised medical records into outpatient-style diagnostic dialogues following the annotation scheme in Section 3.3 and the handbook workflow. For each of the six target diseases, 20 expert-written records are created, yielding 120 seed cases.

Second, we use o3 with one-shot prompting to generate 80 augmented dialogues per disease, resulting in 480 synthetic cases. Following expert advice, selected disease-defining features are removed from some prompts to create more difficult differential-diagnosis cases.

Third, all generated dialogues undergo quality control. Two researchers independently review each dialogue for information completeness, temporal coherence, premature diagnosis leakage, label consistency, and evidential scope. Disagreements are resolved by a senior vascular surgeon. Further details are provided in Appendix A.

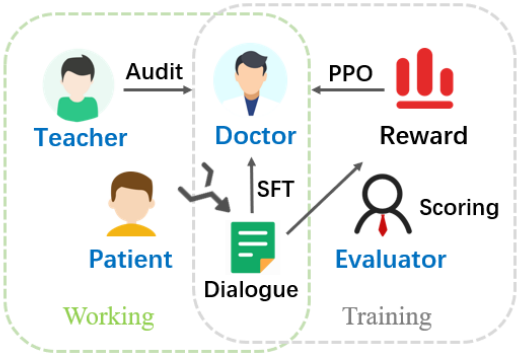


Figure 1: Overview of the proposed multi-agent diagnostic framework, illustrating the interaction between the Doctor, Patient, Teacher, and Evaluator agents.

4 Method

4.1 Process-Aligned Diagnostic Loop

We model outpatient diagnosis as a process-aligned, multi-turn hypothetico-deductive loop. At each consultation turn, the model performs four sequential reasoning stages: Summary, Inference, Question planning, and Diagnosis. The first three stages are repeated across turns to refine the differential diagnosis and reduce uncertainty, while the diagnosis stage is executed only when sufficient evidence has been collected.

More concretely, the summary stage updates a structured representation of the patient's current condition based on the dialogue history. The inference stage revises the differential diagnosis and the relative likelihood of candidate diseases. The question planning stage generates targeted follow-up questions intended to resolve remaining ambiguity. Finally, when the model determines that diagnostic uncertainty has been reduced sufficiently, it enters the diagnosis stage and produces the final diagnostic decision together with a disease-trend report. This staged formulation is also consistent with the annotation structure of LEA-Dialog, examples of which are provided in Appendix C.3, Appendix C.5, and Appendix C.6, and supports supervision for both intermediate reasoning and final outcomes.

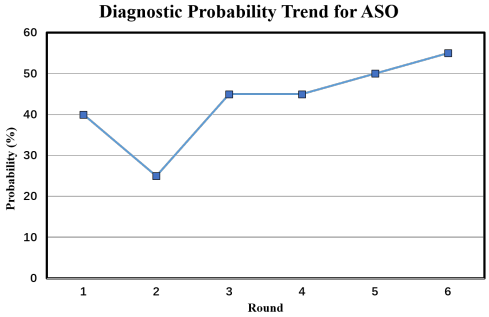
4.2 Multi-Agent Diagnostic Framework

Building on this diagnostic loop, we propose a multi-agent framework consisting of four roles: a Doctor agent, a Patient agent, a Teacher agent, and an Evaluator agent, shown in Figure 1. The Doctor agent is the trainable policy and is responsible for conducting the consultation. The Patient agent simulates the patient side of the dialogue by responding according to the case history. The Teacher agent acts as a clinical instructor who compares the Doctor's behaviour against the guideline file and provides structured feedback. The Evaluator agent converts the consultation trace into optimisation signals for training. The prompts and outputs example for each agent are provided in Appendix C.

The Doctor agent follows the four-stage reasoning structure at each turn. By explicitly separating summarisation, differential refinement, follow-up questioning, and diagnosis, the Doctor agent produces outputs that are easier to audit than a single free-form response. In the question-planning stage, particular emphasis is placed on clarifying ambiguous or potentially incomplete patient descriptions, reflecting the uncertainty commonly encountered in outpatient consultations.

The Patient agent plays the role of the patient and answers questions using the diagnostic history provided in the dataset. This design allows us to simulate multi-turn consultations in a controlled setting while keeping the dialogue grounded in case-specific evidence.

The Teacher agent evaluates the Doctor's reasoning process after each turn by cross-referencing it with the companion guideline file. It scores the appropriateness of the Doctor's actions, identifies omitted or mistimed questions, and provides strategic feedback such as revisiting a previous step or proceeding to the next stage. After the final diagnosis, the Teacher agent also generates a structured summary of the diagnostic trajectory, including the historical evolution of disease-probability assessments. This report provides a turn-level reference for subsequent training and evaluation.



- (1) 40% in Round 1 based only on sudden exertional pain;
- (2) Lowered to 25% in Round 2 due to lack of classic 5 Ps;
- (3) Raised to 45% in Round 3 with typical claudication (100m pain, relief with rest) and no venous signs;
- (4) Kept at 45% in Round 4 with sedentary lifestyle and cessation of anticoagulation but no strong trigger;
- (5) Raised to 50% in Round5 due to prior DVT with anticoagulation stop, suggesting a prothrombotic milieu;
- (6) Final 55% since still no swelling/warmth and arterial risk predominates.

Figure 2: Visualisation of diagnostic probability trends across rounds. The figure shows the probability assigned by the doctor agent to ASO at the end of each round, with the corresponding rationale. The full report includes probability trends and explanations for all candidate diseases.

4.3 Teacher-Guided Diagnostic Trajectory Analysis Report

A key component of our framework is the structured diagnostic trace produced by the Teacher agent. Rather than treating the consultation as an unstructured sequence of responses, we represent it as a partial diagnostic trend grounded in clinical guideline nodes shown in Figure 2. Each consultation turn contributes to this trend by adding or revising evidence, narrowing the differential diagnosis, and updating probability trends over candidate diseases. The final report, therefore, captures not only the final diagnosis but also whether the model followed an appropriate diagnostic trajectory.

This representation serves two purposes. First, it improves interpretability by making intermediate decisions easier to inspect. Second, it enables turn-level supervision beyond final-answer accuracy. In particular, we use the Teacher-generated report to assess whether the Doctor agent followed the handbook guidance, whether probability trends changed in clinically plausible directions, and whether unsupported claims were introduced during reasoning.

4.4 Turn-Level optimisation

To optimise the Doctor agent for both diagnostic correctness and process stability, we define the total reward as a weighted combination of turn-level rewards and a final reward:

$$G = \sum_{t=1}^{T} R_t^{\text{inst}} + R^{\text{final}}$$

where R_t^{inst} is the immediate reward at turn t,R^{final} is the final reward for the completed consultation. Intuitively, the turn-level reward encourages the Doctor agent to follow a clinically coherent diagnostic trajectory by rewarding appropriate guideline-node expansion and correct probability-trend updates while penalising unsupported claims. The final reward evaluates the overall diagnostic quality, including final diagnosis correctness and trajectory-level similarity to the reference diagnostic tree. Full details of the reward design are deferred to Appendix D.

5 Experimental Setup

5.1 Models

We evaluate both online and offline LLMs. For the Doctor agent, we use two online models, GPT-4o and o3, and two offline models, Qwen3-8B and Qwen3-14B. For the two offline models, we further evaluate their framework-trained versions after supervised fine-tuning and PPO, denoted as Qwen3-8B-T and Qwen3-14B-T, respectively. This setup allows us to test whether the proposed framework improves not only strong proprietary models but also smaller deployable open-weight models, and whether process-aligned optimisation can reduce the gap between offline and online systems. Unless otherwise specified, GPT-4o is used for the Patient, Teacher, and Evaluator roles.

5.2 Training and Comparison Settings

The Doctor agent is optimised using a hybrid PPO+SFT strategy; full details of the training framework are provided in Appendix E. During training, the Doctor interacts with the Patient agent in a multi-round consultation setting. Based on the interaction history, the Teacher agent audits guideline-node coverage, probability-trend updates, and unsupported claims, and provides structured process-level feedback. The Evaluator agent then converts these signals, together with the final diagnostic outcome, into training rewards. To maintain

language fluency and structural consistency, we interleave reinforcement learning with supervised fine-tuning, performing one SFT round after every five PPO rounds.

We compare each framework-enhanced Doctor model with its corresponding base model under direct prompting. We further conduct ablation studies on three key components of the framework: the companion guideline file, the Teacher agent, and the turn-level optimisation procedure, to isolate the contributions of process alignment, structured feedback, and reward-based optimisation.

5.3 Evaluation Metrics

We evaluate the framework from four perspectives: final diagnostic performance, Teacher reliability, repeated-run stability, and process-level error patterns.

For final diagnostic performance, we report final accuracy. For each case, the model outputs a probability distribution over the candidate diseases. The disease with the highest final probability is taken as the prediction and compared with the reference clinical diagnosis. A prediction is counted as correct if the predicted and reference labels refer to the same disease category; this matching is manually verified to avoid surface-form variation. Formally,

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i],$$

where N is the number of evaluated cases, $\hat{y}_i$ is the manually verified top-probability diagnosis, and y_i is the reference diagnosis.

For Teacher reliability, we evaluate three binary process-level error types: record-inconsistent summary, reasoning non-adherence, and questioning non-adherence. These correspond to contradiction or omission in the summary, deviation from handbook-defined diagnostic logic, and failure to ask guideline-relevant or discriminative follow-up questions. We compare the Teacher's judgements with expert annotations using confusion matrices. The validation sample contains 102 dialogue histories, with 17 instances per disease, each independently assessed by three clinical experts.

For repeated-run stability, we report Rate-n. For each test case, the model is run multiple times, and a run is counted as stable up to turn n if the trend direction of the correct disease matches the majority trend direction at each of the first n turns.

Model	Base	Multi Only	HB Only	Multi+HB
Qwen3-8B	19.8	27.2	62.1	89.7
Qwen3-8B-T	23.2	27.3	65.4	90.3
Qwen3-14B	23.5	31.0	69.3	92.3
Qwen3-14B-T	25.7	31.3	71.1	94.0
GPT-4o	87.6	89.3	90.9	96.0
o3	93.1	93.3	96.9	97.7

Table 2: Performance comparison (Acc %) under different settings on the held-out test set ($N = 60$, 10 cases per disease). Qwen3-8B-T and Qwen3-14B-T denote the models after PPO+SFT training. Multi refers to the multi-step reasoning scheme, while HB refers to the use of the handbook to provide diagnostic guidance and supplementary knowledge to the model.

Finally, we analyse misdiagnosed trajectories to identify the first clinically meaningful deviation and its dominant error type.

6 Results

6.1 Main Results

Table 2 summarises the main diagnostic results. Overall, the proposed framework with the handbook improves performance across both online and offline models, with especially strong gains for the smaller offline models. In particular, framework-enhanced Qwen3-14B-T with handbook improves average diagnostic accuracy by 68.3% and surpasses the original o3 model by 0.9%, showing that process-aligned optimisation and refined diagnosis handbook can substantially narrow the gap between deployable offline models and strong online models. For o3, the framework still yields measurable benefits, improving average diagnostic accuracy by 4.6% and reducing the misdiagnosis rate from 6.9% to 2.3%. The handbook is a primary source of the observed performance gains. In the handbook ablation study, access to the handbook substantially improves diagnostic accuracy for offline models and also yields a smaller but consistent improvement for online models. Notably, this benefit becomes larger after applying the multi-step framework, suggesting that the proposed multi-stage diagnostic procedure is more effective at incorporating external clinical knowledge. In contrast, when the underlying diagnostic logic is unchanged, reinforcement learning alone provides only limited gains in accuracy. These results indicate that, for final diagnostic performance, the effective use of external clinical knowledge is more critical than policy optimisation alone. This finding highlights the importance of knowledge utilisation,

rather than optimisation alone, in process-aligned diagnostic systems.

6.2 Teacher Reliability and Ablation

We next provide a more detailed analysis of the Teacher agent's performance. Specifically, we first evaluate whether the Teacher can correctly identify unsupported claims produced by the Doctor agent. Naturally, given the three stage labels in each diagnostic round—Summary, Reasoning, and Question Planning—we categorize such unsupported behaviours into three corresponding types: Record-inconsistent summary, where the Doctor's summary is inconsistent with the patient record or omits key clinical information; Reasoning non-adherence, where the Doctor's reasoning deviates from the guideline; and Questioning non-adherence, where the Doctor's planned follow-up questions do not conform to the guideline requirements.

The confusion matrix for Questioning non-adherence is shown in Figure 3, while the corresponding results for the other two criteria are provided in Appendix F. The confusion matrices across the three unsupported-claim criteria show that the Teacher's behaviour varies by criterion. For record-inconsistent summaries and reasoning guideline non-adherence, the Teacher is relatively conservative, producing few false positives but missing a non-trivial portion of rare errors. For questioning guideline non-adherence, it is more sensitive, identifying more problematic questioning steps at the cost of increased false positives and additional re-questioning. These results suggest that the Teacher does not apply a uniform decision rule, but instead exhibits a criterion-dependent safety–efficiency trade-off.

Teacher ablation further shows that removing the Teacher yields only a modest change in overall diagnostic accuracy, but substantially weakens record consistency, guideline questioning adherence, and guideline reasoning adherence. This finding indicates that the Teacher's main contribution is not simply improving final accuracy, but regularising the diagnostic process toward handbook-consistent and more trustworthy behaviour.

6.3 Stability Analysis

In our experiments, we observed that when the model repeatedly diagnoses the same patient, the diagnostic trajectories tend to remain highly similar across successful runs, as illustrated by Test 1–3 in Figure 4, whereas failed runs (Test 4–5) often

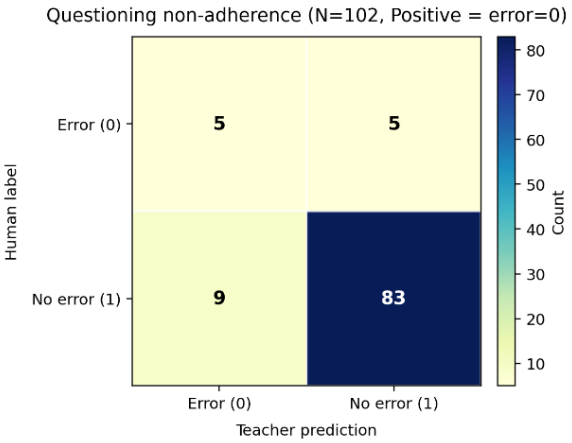


Figure 3: Confusion matrices for questioning non-adherence. In total, we sampled 102 dialogue histories, with 17 instances per disease. Each history was independently assessed by three experts for the three types of unsupported behaviours. Histories without unsupported behaviour were labelled as Positive (1), whereas those containing unsupported behaviour were labelled as Negative (0).

follow different trajectories. This divergence arises because the model cannot perfectly adhere to the guidelines in every run. Under the assumption that the diagnostic guideline is of sufficiently high quality, the stability of the model's outputs becomes an important determinant of diagnostic accuracy. Motivated by this observation, we introduce Rate-n to evaluate output stability.

Rate-n measures whether, across 10 repeated runs on the same case, the model maintains a similar diagnostic trajectory for the correct disease within the first n rounds. Here, a *similar trajectory* refers to the majority trend direction observed across runs, that is, whether the probability of the correct disease increases or decreases at each turn. Using Figure 4 as an example, among 10 repeated runs, the model correctly identified the possibility of ASO in only 5 runs at Round 1 (namely, Test1–5), yielding a Rate-1 of 50%. At Round 2, Test1, Test2, Test3, and Test5 shared the same trend for ASO (a decrease), which constitutes the majority pattern, yielding a Rate-2 of 40%. By the same logic, Rate-3 is 30%.

It is worth noting that a decrease in the probability trend of the correct disease at a particular round does not necessarily indicate an error in either the model or the guideline. Rather, it may reflect that the newly collected evidence at that stage also supports other competing diagnoses, which is common in real clinical reasoning.

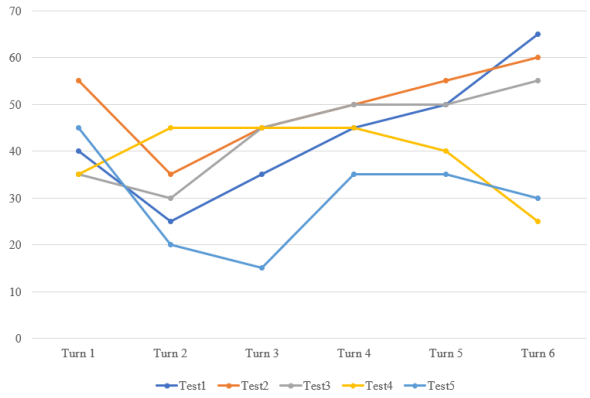


Figure 4: Evolution of the model's inferred probability for ASO across sequential rounds in an ASO clinical case (%). Among them, Tests 1–3 correspond to successful diagnoses and show similar diagnostic trends, whereas Tests 4 and 5 correspond to failed diagnoses and follow trajectories that differ from the others.

Model	Rate-1	Rate-2	Rate-3
Qwen3-8B	91.7	76.7	58.3
Qwen3-8B-T	91.7	81.7	61.7
Qwen3-14B	93.3	80.0	61.7
Qwen3-14B-T	95.0	83.3	66.7
GPT-4o	96.7	76.7	45.0
o3	98.3	78.3	41.7

Table 3: Stable diagnostic rates across sequential turns for six distinct disease categories

As shown in Table 3, across models, large online models show a sharper decline in Rate-n after the first turn, whereas small offline models exhibit a smaller drop. The PPO-trained offline models are the most stable overall. In particular, Qwen3-14B-T outperforms o3 by 5% from Turn 2 onward and by 25% from Turn 3 onward, showing that process-aligned RL can substantially reduce reasoning drift in multi-turn diagnosis.

This result shows that stronger base capability does not necessarily translate into more stable diagnostic trajectories. In our setting, reinforcement learning (PPO) acts as a regulariser that keeps the model more tightly aligned with the structured knowledge source, leading to more consistent reasoning across repeated trials.

6.4 Error Pattern Analysis

We further conduct a retrospective error-pattern analysis on 60 misdiagnosed trajectories to identify both the earliest clinically meaningful deviation and its primary error type. We annotate three error categories:
(1) Summary error, where the model distorts the patient summary or omits key symptoms;
(2) Reasoning error, where the model fails to follow the handbook-defined differential diagnostic logic;
(3) Question-planning error, where the model fails to ask the most discriminative follow-up question at a critical decision point.

As shown in Table 4, the dominant failure mode is question-planning error, which accounts for 53.3% of all cases. This indicates that the main bottleneck is often not missing medical knowledge, but failure to ask the most discriminative next question at critical branching points. Reasoning errors are the second most frequent category (40.0%), while summary errors are less common but tend to occur earlier and distort downstream reasoning.

Errors also emerge early: 71.7% of trajectories show their first meaningful deviation within the first two turns. Summary errors tend to concentrate in Turn 1, question-planning errors in Turn 2, and reasoning errors more often from Turn 2 onward. These findings suggest that misdiagnosis rarely results from a single final-turn mistake, but instead develops progressively from earlier process-level deviations.

Error Type	Number	Rate
Question Planning Error	32	53.3%
Reasoning Error	24	40.0%
Summary Error	4	6.7%
Total	60	100.0%

Table 4: Distribution of process-level error types (N=60)

7 Conclusion

We introduced LEA-Dialog and a process-aligned agentic framework for multi-turn diagnosis of lower-extremity arteriovenous diseases. The results show that evaluating only final accuracy is insufficient: guideline-grounded reasoning, process-level errors, and repeated-run stability are also important for assessing diagnostic dialogue systems. Structured clinical guidance and process-level optimisation improve performance and reduce reasoning drift, particularly for smaller offline models. Future work will broaden disease coverage and include more real clinical interactions.

8 Limitations

This study has several limitations. LEA-Dialog focuses on a narrow but clinically coherent group of six lower-extremity arteriovenous diseases. This enables controlled analysis of diagnostic trajectories among closely related vascular conditions, but limits claims about generalisation to broader medical domains. In addition, part of the dataset is synthetically expanded from expert-written seed cases, which may introduce distributional bias and make the dialogues more regular than real outpatient interactions. We therefore view LEA-Dialog as a controlled, domain-specific testbed for process-aligned diagnostic reasoning rather than a fully representative corpus of clinical consultations. The simulated multi-agent setup and the use of GPT-4o for the Patient, Teacher, and Evaluator roles may further introduce an LLM echo-chamber effect. We mitigate this risk by grounding patient responses in case records, using a guideline-based handbook, and validating a subset of Teacher judgements against expert annotations. Due to clinical data governance constraints, the original expert-derived records cannot be fully released. We plan to release the synthetic dialogues, annotation schema, prompts, handbook structure, and evaluation scripts where permitted, subject to institutional approval. Future work should broaden disease coverage, include more real clinical interactions, and validate the framework through independent clinician-in-the-loop studies.

Acknowledgements

This work was partly supported by EPSRC via Responsible AI UK (grant number EP/Y009800/1, project KP0016 AdSoLve) and the Slovenian Research Agency (ARRS) under core research programme Knowledge Technologies (P2-0103).

References

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. 2025. AI Hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. MS-Diagnosis: An EMR-based dataset for clinical multi-step diagnosis. *arXiv preprint arXiv:2408.10039*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model Meta AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6).

Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. 2024. MedBench: A comprehensive, standardised, and reliable benchmarking system for evaluating Chinese medical large language models. *Big Data Mining and Analytics*, 7(4):1116–1128.

Xinyi Liu, Dachun Sun, Yi Fung, Dilek Hakkani-Tur, and Tarek F Abdelzaher. 2025. DocCHA: Towards LLM-augmented interactive online diagnosis system. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 609–619.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, Ming Zhang, Fangzhou Li, Xiaowei Ma, Yue Shen, Jinjie Gu, Wei Xue, and Yiran Huang. 2024. RJUA-QA: A comprehensive QA dataset for urology.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. 2024. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI*, 3:e53207.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. *Nature Medicine*, 31(3):943–950.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, {S. Sara} Mahdavi, Christopher Semturs, and 7 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Iris Vessey and Dennis Galletta. 1991. Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, 2(1):63–84.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen,

Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6796–6814.

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. DrHouse: An LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29.

A Detailed Case Collection Procedure

A.1 Study Design and Data Sources

This study is a single-center retrospective study. Cases were collected from routine clinical care at a tertiary hospital, including the electronic medical record system, laboratory information system, imaging reports, surgical records, and discharge summaries. The study population included patients with DVT, PTS, lower-extremity varicose veins (VV), ASO, AT, and Buerger's disease. As this study is based on routinely collected clinical data and aims to construct an observational research dataset, the data reporting followed the STROBE, RECORD, and TRIPOD-LLM guidelines.

A.2 Ethical Compliance and De-identification

The study was approved by the institutional ethics committee and conducted in accordance with the Declaration of Helsinki, which explicitly applies to medical research involving identifiable human materials or data. All patient information was de-identified before being entered into the research database, with names, national ID numbers, hospital admission numbers, telephone numbers, detailed addresses, and other direct identifiers removed.

A.3 Inclusion and Exclusion Criteria

Included cases had to meet two criteria: first, a clearly established final clinical diagnosis; second, at least one core piece of evidence supporting that diagnosis, such as venous ultrasound findings or anticoagulation records for DVT, Doppler ultrasound for varicose veins, CTA/DSA or intraoperative records for AT, and smoking history together with characteristic imaging or pathological evidence for Buerger's disease.

Exclusion criteria included missing key fields, repeated hospitalisations that could not be distinguished as independent cases, insufficient diagnostic evidence, severely incomplete text, or cases in which reliable de-identification could not be achieved.

A.4 Structured Extraction Fields

Each case was converted into a structured record using a unified schema, with source materials organised in the following order:

1. **Demographic information:** age, sex, and admission date.
2. **Chief complaint:** the most prominent presenting symptom and its duration at the first visit.
3. **History of present illness:** onset pattern, progression speed, triggers, affected site, associated symptoms, and prior management.
4. **Past medical history:** prior thrombosis, varicose vein history, PAD/cardiovascular disease history, diabetes, hypertension, chronic kidney disease, malignancy, surgical history, and prior anticoagulant/antiplatelet use.
5. **Personal history:** smoking history, which was especially important to preserve for Buerger's disease cases.
6. **Physical examination:** limb swelling, tenderness, skin temperature, skin colour, ulceration/gangrene, superficial venous dilation, and dorsalis pedis/posterior tibial pulse findings.
7. **Auxiliary examinations:** D-dimer, venous/arteriovenous ultrasound, ABI/TBI, CTA/MRA/DSA, and intraoperative findings.
8. **Diagnostic information:** preliminary diagnosis, final diagnosis, subtype/stage, or relevant severity score.
9. **Treatment information:** medication, intervention, surgery, and perioperative management.

A.5 Key Information Extraction and Standardisation

Three researchers with vascular surgery backgrounds developed the extraction dictionary and mapping rules in advance to standardise symptoms, examinations, anatomical sites, diagnostic terms, and treatment modalities. For example, expressions such as "lower-extremity swelling and pain" and "calf swelling with pain" were consolidated into the same symptom category, while "superficial femoral artery occlusion" and "SFA occlusion" were mapped to the same lesion-site category.

A.6 Data Construction and Augmentation

Three physicians with vascular surgery backgrounds expanded the highly standardised and structured medical records into diagnostic narratives based on the diagnostic workflow specified in the handbook and their real clinical experience. The expanded records were then presented in the form of outpatient-style dialogues, which served as the seed set. For each disease, 20 diagnostic records were created in this way. Subsequently, o3 was prompted with a one-shot setup based on the seed set to simulate and generate additional virtual diagnostic dialogues, producing 80 dialogues per disease as the augmented set.

A.7 Quality Control and Dataset Split

All generated dialogues were independently reviewed by two researchers. The review focused on whether the information was complete, whether the temporal order was coherent, whether the dialogue leaked the final diagnosis prematurely, whether the diagnostic label was consistent with the original medical record, and whether the management suggestions remained within the evidential scope of the source case. Any disagreements were resolved by a senior vascular surgeon.

After quality control, the finalised cases were used for dataset construction and subsequent train–validation–test splitting.

B Handbook Development Plan and Reference Guidelines

B.1 Handbook Development Strategy

The handbook used in this study was developed for the diagnostic setting of lower-extremity arteriovenous diseases. It covered six target diseases: deep vein thrombosis (DVT), post-thrombotic syndrome (PTS), lower-extremity varicose veins(VV), arteriosclerosis obliterans (ASO), AT, and thromboangiitis obliterans (Buerger disease).

The primary evidence sources for the handbook were recent international clinical practice guidelines, diagnostic consensus statements, and other normative clinical documents. Specifically, the DVT and PTS sections were mainly based on the ESVS 2021 venous thrombosis guideline, the CHEST guideline for venous thromboembolism, and NICE NG158 (*Venous thromboembolic diseases: diagnosis, management and thrombophilia testing*). The chronic venous disease and varicose vein sections were mainly based on the ESVS 2022 chronic venous disease guideline, the SVS/AVF/AVLS guideline for varicose veins, and NICE CG168 (*Varicose veins: diagnosis and management*). The ASO and AT sections were mainly based on the 2024 ACC/AHA guideline for lower-extremity peripheral artery disease, the 2024 ESC guideline for peripheral arteriovenous and aortic diseases, and the ESVS 2020 acute limb ischemia guideline. The Buerger disease section was mainly based on the 2023 international diagnostic consensus, supplemented by recent review and treatment evidence.

To ensure consistency between the knowledge base and real-world clinical reasoning, the handbook was uniformly organised around a six-step consultation workflow. This design allowed each knowledge entry to align with the sequential logic of outpatient vascular diagnosis.

B.2 Six-Step Consultation Workflow

The handbook was structured according to the following six-step consultation process:

1. **Comprehensive symptom inquiry.** The clinician should thoroughly review and further probe the patient's lower-extremity symptoms to avoid missing manifestations with important differential diagnostic value.

2. **Past medical history inquiry.** This step includes questions about comorbidities associated with lower-extremity vascular disease, such as hypertension, diabetes, and coronary artery disease, as well as chronic infectious diseases relevant to perioperative assessment, such as tuberculosis and hepatitis B. Previous thrombotic history should also be explicitly assessed.

3. **Lifestyle and demographic inquiry.** The

clinician should ask about smoking and alcohol use, lifestyle patterns related to prolonged sitting or standing that may contribute to varicose veins, and demographic features associated with specific diseases, such as the typical profile of Buerger's disease patients (e.g., young or middle-aged men). Age and sex should also be collected at this stage if they were not obtained in the first turn, since they are useful for distinguishing age-related vascular conditions.

4. **Previous surgery and treatment history.** This step covers prior surgical history and previous disease management measures, including medication, physical therapy, and other interventions.

5. **Structured summary of effective information.** The clinician should summarise all diagnostically useful information obtained from the previous rounds of questioning.

6. **Final diagnostic output.** The diagnosis should be presented in the format `[Disease A: probability], [Disease B: probability], . . .` until all candidate diseases and their estimated probabilities have been listed. The diagnostic output should terminate with `F/`.

B.3 Rationale for This Structure

This workflow-based design was intended to make the handbook not only a static collection of disease knowledge but also a process-aligned clinical reasoning resource. Instead of presenting isolated disease descriptions, the handbook encoded what information should be asked, when it should be asked, how previously collected evidence should be summarised, and how differential diagnoses should be updated before the final decision. In this way, the handbook better matched the iterative and hypothetico-deductive nature of real outpatient diagnosis.

B.4 Reference Guidelines Used for Handbook Development

The following references were used as the main evidence base for handbook construction:

1. Gallifant J, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* 31, 60–69 (2025).
2. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335, 806–808 (2007).
3. Benchimol EI, et al. The Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) statement. *PLoS Medicine* 12, e1001885 (2015).
4. World Medical Association. Declaration of Helsinki: Ethical principles for medical research involving human participants. *JAMA* 333, 71–74 (2025).
5. Sounderajah V, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nature Medicine* 31, 3283–3289 (2025).
6. Kakkos SK, et al. Editor's Choice – European Society for Vascular Surgery (ESVS) 2021 Clinical Practice Guidelines on the Management of Venous Thrombosis. *European Journal of Vascular and Endovascular Surgery* 61, 9–82 (2021).
7. Stevens SM, et al. Antithrombotic Therapy for VTE Disease: Second Update of the CHEST Guideline and Expert Panel Report. *Chest* 160, e545–e608 (2021).
8. De Maeseneer MG, et al. Editor's Choice – European Society for Vascular Surgery (ESVS) 2022 Clinical Practice Guidelines on the Management of Chronic Venous Disease of the Lower Limbs. *European Journal of Vascular and Endovascular Surgery* 63, 184–267 (2022).
9. Gloviczki P, et al. The 2023 Society for Vascular Surgery, American Venous Forum, and American Vein and Lymphatic Society clinical practice guidelines for the management of varicose veins of the lower extremities. Part II: Endorsed by the Society of Interventional Radiology and the Society for Vascular Medicine. *Journal of Vascular Surgery: Venous and Lymphatic Disorders* 12, 101670 (2024).
10. Gornik HL, et al. 2024 ACC/AHA/AACVPR/APMA/ABC/SCAI/SVM/SVN/SVS/SIR/VESS Guideline for the Management of

Lower Extremity Peripheral Artery Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* 149, e1313–e1410 (2024).

11. Mazzolai L, et al. 2024 ESC Guidelines for the management of peripheral arterial and aortic diseases. *European Heart Journal* 45, 3538–3700 (2024).

12. Björck M, et al. Editor's Choice – European Society for Vascular Surgery (ESVS) 2020 Clinical Practice Guidelines on the Management of Acute Limb Ischaemia. *European Journal of Vascular and Endovascular Surgery* 59, 173–218 (2020).

13. Fazeli B, et al. Diagnostic criteria for Buerger's disease: International Consensus of VAS – European Independent Foundation in Angiology/Vascular Medicine. *International Angiology* 42, 396–401 (2023).

14. Uyanık SA, et al. Endovascular Treatment of Critical Limb Ischemia in Buerger Disease (Thromboangiitis Obliterans) With Midterm Follow-Up: A Viable Option When Bypass Surgery Is Not Feasible. *AJR American Journal of Roentgenology* 216, 421–427 (2021).

C Prompt details and output examples

C.1 Doctor-Agent Prompt

You are a physician specialising in lower-limb arteriovenous and venous diseases. In each round, you question the patient according to the uploaded diagnostic guideline and report your reasoning and question-planning process to your mentor, conducting the diagnosis under the mentor's supervision. You must strictly follow the six-step diagnostic process described in the handbook — do not change the order of questioning or skip any step arbitrarily. Depending on who is replying to you, you will perform different steps.

After each round of the patient's response, proceed as follows: Confirm the current diagnostic step according to the guideline. Execute the content required for the current step.

The execution of each diagnostic step should include:

Summary: summarise the patient's current condition using professional terminology based on their latest response.

Reasoning: analyse which parts of your previous reasoning may have been incorrect according to the new response, revise them, and analyse the possible diseases in the format [Disease A: possible reason A], [Disease B: possible reason B], ... to present the updated reasoning results.

Diagnostic Judgment: Determine whether the current information (excluding imaging data) is sufficient to begin diagnosis. – If sufficient, proceed directly to Step 5 (Diagnosis). – If insufficient, proceed to Step 4 (Question Planning).

Question Planning: Evaluate whether the patient's answers contain vague descriptions that could significantly affect diagnosis. – If yes, list them in the format [Vague description A: possible inaccuracy 1], [Vague description B: possible inaccuracy 2], ... then draft clarification questions and skip Step 5. – If no ambiguity, formulate the next set of questions according to the current diagnostic step and skip Step 5.

Diagnosis: Output all possible diseases and their probabilities in the format [Disease A: probability], [Disease B: probability], ... until all have been listed. Refusal to provide a diagnosis is not allowed. End the diagnostic output with "F/".

After each round of the mentor's response: – If the mentor instructs you to continue questioning, proceed to ask the patient the questions you drafted in the previous round. – If the mentor instructs you to reconsider, repeat the previous step from the start.

The first round of dialogue should begin with: "Hello, what can I help you with today?"

C.2 Patient-Agent Prompt

You are a patient. Read the uploaded historical diagnostic record and answer the doctor's questions strictly according to the information in that record. You must not fabricate any symptoms or conditions that are not mentioned in the case history. When describing your symptoms, speak in a natural, conversational tone, but ensure your answers are as complete and consistent with the case record as possible. In the closed question–answer dialogue, if the doctor asks about any symptom or manifestation not included in the record, always answer that you are not sure, and do not invent any new symptoms or medical history. Each time you reply, provide your entire answer in one paragraph. Use the first paragraph of the patient's description from the diagnostic record as your opening statement in the conversation with the doctor.

C.3 Doctor-Patient output example

Doctor A: Hello, how can I help you today?
Patient A: I am an 84-year-old man . . .

Doctor A: [**Current Step**]
Step 2. Review and clarify the patient's lower-limb symptoms.

[1. Summary]
Male, 84. Sudden onset of right-lower-limb pain 4 days ago, worse with walking, relieved by rest. ED initially suspected DVT . . .

[2. Reasoning Revision]
[Acute limb ischemia: could be in-situ arterial thrombosis or embolism; acute onset + exertional pain suggests ischemic pain.]
[Deep vein thrombosis (DVT): ED's first consideration; need to confirm swelling, tenderness, increased skin warmth, etc.]

[3. Diagnostic Judgment]
Current information is insufficient to start diagnosis; we need details on accompanying symptoms/signs and vascular pulses.

[4. Question Plan]
Ambiguous expressions
[The nature and location of "pain" are unclear: cannot distinguish ischemic from venous aching.]

Please answer the following one by one:
1) Is the pain in the whole right leg, or specifically in the thigh, calf, or foot? Is it aching, dull, stabbing, or cutting?
2) Over the past 4 days, has the right lower limb become clearly swollen? Any difference in girth compared with the left?

[Possible Diseases and Probabilities (Initial)]
[Acute lower-limb arterial ischemia/embolism: 40%]
[Deep vein thrombosis (DVT): 35%]
[Acute worsening of lower-limb atherosclerotic occlusive disease: 25%]

C.4 Teacher-Agent Prompt

You are Doctor A's mentor. You may only communicate with Doctor A and must not speak directly with Patient A.

Each dialogue round includes either a Suggestion Module or a Report Module. Determine whether Patient A has triggered the diagnostic module (ends with "F/"). – If not triggered, enter the Suggestion Module. – If triggered, enter the Report Module.

Suggestion Module

You should read the uploaded treatment guideline and assess whether Doctor A has: (1) Record-consistent summary: Provide an accurate summary of the patient's status. (2) Guideline non-adherence during reasoning: Followed the correct diagnostic logic from the handbook. (3) Guideline non-adherence during questioning: Followed the correct questioning sequence from the handbook.

For each of the three criteria above, we assign a binary score of 0 or 1: 1 indicates the model makes no error, and 0 indicates an error occurs. If the total score is below 3, we prompt the model to re-initiate the questioning round; otherwise, we instruct it to continue the diagnostic process and provide recommendations for the next-step clarification questions.

Note: Doctor A's first round of dialogue does not require analysis—simply tell Doctor A to continue.

Report Module

Summarise every change in the probability of each possible disease throughout Doctor A's entire diagnostic process, along with the reason for each change. Use the following format: [Disease A: probability1–probability2–probability3. . . , Reason: . . .]

End the report with "E/"

C.5 Teacher-Agent Output example (Suggestion)

Coordinator Agent: *Score*

Record-inconsistent summary: 1

Doctor A accurately restated the patient's most recent response (54-year-old male; former smoker; no alcohol use; no prolonged sitting/standing).

Guideline adherence during reasoning: 1

Checkpoint 4 has been completed, and the dialogue correctly proceeds to the inquiries required for Checkpoint 5; the diagnostic logic is consistent with the handbook.

Guideline adherence during questioning: 1

The two questions asked (history of lower-limb-related surgeries and long-term medication use) fully meet the requirements of Checkpoint 5.

Total: 3/3 — The diagnostic workflow may continue.

Improvements and next-step recommendations

After the patient answers Checkpoint 5, please immediately execute the Checkpoint 6 "double-summary" requirement.

Key information reminders

The patient's ultrasound has shown "multiple deep vein thromboses in the right lower limb." If there is no prior DVT history but the course has lasted 2 months, consider subacute DVT and early post-thrombotic syndrome (PTS) as the top two differentials.

C.6 Teacher-Agent Output example (Report)

Coordinator Agent:

[Right-lower-limb subacute arterial ischemia / ASO acute exacerbation: 40–25–45–45–50–55.]

Reasons: Round 1 gave 40% based only on "sudden exertional pain"; Round 2 lowered to 25% due to lack of classic 5 Ps; Round 3 raised to 45% with typical claudication (100 m pain, relief with rest) and no venous signs; Round 4 kept at 45% with a sedentary lifestyle and cessation of anticoagulation, but no strong trigger; Round 5 raised to 50% due to prior DVT with anticoagulation stop, suggesting a prothrombotic milieu; Final 55% since still no swelling/warmth, and arterial risk predominates.

[Recurrent DVT: 35–15–10–20–20–20.]

Reasons: Initially 35% for "ED suspected DVT + prior DVT"; Round 2 dropped to 15% with no swelling/tenderness; Round 3 down to 10% with continued lack of venous support; After learning of stopping anticoagulation and sedentary lifestyle, the rate rose to 20% were maintained.

[Distal arterial embolism: 40–25–45–20–15–15.]

Reasons: Began with 40% for all acute ischemia; Without cardiac source, lowered to 25%; When not yet excluded, considered alongside in-situ thrombosis back to 45%; With no heart source, down to 20%, and then maintained at 15%.

[Chronic ASO only (stable): 20–45–35–10–10–8.]

Reasons: Claudication with rest relief once raised it to 45%; No previous claudication and sudden onset, then reduced; Sedentary lifestyle, but no long history of claudication, dropped it to 10%; Finally 8% due to acute characteristics and stopping anticoagulation.

[Venous insufficiency/varicosities: 5–10–5–0–0–0.]

Reasons: Briefly raised to 10% in Round 2 on "nocturnal cramps/heaviness" query, then fell to zero due to continuous lack of supporting signs.

E/

D Reward function design

The immediate reward score R_t^{inst} integrates node additions, correctness of probability trend assessments, and unsupported-claims penalties, while the final reward score R^{final} combines overall accuracy with the longest common subsequence (LCS) similarity of the diagnostic tree:

$$G = \sum_{t=1}^{T} R_t^{\text{inst}} + R^{\text{final}} \tag{1}$$

Specifically,

$$G = \sum_{t=1}^{T} \big(\alpha\,\Delta\text{Node}_t + \beta\,\text{Trend}_t - \gamma\,\text{Halluc}_t\big) + \lambda\,\text{Acc} + \mu\,\text{LCS} \tag{2}$$

Above all, $\text{Acc} \in \{0, 1\}$, $\text{LCS} \in [0, 1]$, $\alpha = 0.3$, $\beta = 0.5$, $\gamma = 0.2$, $\lambda = 0.7$, $\mu = 0.3$.

Node Increment (ΔNode_t). A handbook-node increment of 0.5 is assigned if a newly completed node follows the guideline sequence; otherwise 0.

Trend Assessment (Trend_t).

- Increasing the probability of the correct disease: $+0.5$
- Decreasing the probability of an incorrect disease: $+0.5$
- Decreasing the probability of the correct disease or increasing that of an incorrect disease: -0.5
- No change in trend: 0

Hallucination Penalty (Halluc_t). Presence of unsupported claims incurs a penalty of -0.5 per claim; no penalty is applied if none are present.

Specifically, unsupported clinical claims are detected from three aspects:

- Record-inconsistent patient status summaries
- Guideline non-adherence during reasoning

- Guideline non-adherence during questioning

The Teacher Agent is responsible for executing the detection and assigning scores. A score below a threshold of 3 indicates the presence of unsupported-claims behaviour, which triggers reinitiation of the questioning round.

Final Accuracy (Acc). The accuracy of the final diagnosis is assessed by semantically comparing it with the diagnostic results in the dataset.

E Training Framework Details

This appendix provides additional details of the hybrid PPO+SFT training framework used to optimize the Doctor agent. Our training design aims to improve not only final diagnostic accuracy, but also the quality and stability of the intermediate diagnostic process. To this end, we combine supervised fine-tuning on reward-filtered diagnostic dialogues with reinforcement learning under process-level supervision.

E.1 Training Overview

We adopt a hybrid training strategy that alternates supervised fine-tuning (SFT) and proximal policy optimisation (PPO). The overall training pipeline contains three stages: (1) construction of candidate diagnostic dialogues, (2) reward-based filtering for high-quality trajectory selection, and (3) iterative optimisation of the Doctor agent with interleaved SFT and PPO.

First, candidate diagnostic dialogues are constructed from both expert-authored seed cases and LLM-augmented virtual cases. Each dialogue contains a multi-round outpatient-style consultation trajectory, including the Doctor's intermediate reasoning process and the final diagnostic decision. These candidate dialogues are then scored using our reward framework, which evaluates handbook-node progression, correctness of disease-probability trend updates, and unsupported-claims behaviour. Only high-quality trajectories are retained for subsequent supervised training.

We first apply SFT on the filtered dialogue set to initialise the Doctor agent with stable multi-round interaction patterns, structured reasoning stages, and trend-aware diagnostic updates. Starting from this initialisation, we further optimise the Doctor agent using PPO in an online interaction setting. During PPO training, the Doctor interacts with the Patient agent round by round, while the Teacher agent audits the evolving consultation process and the Evaluator agent converts these signals into training rewards. To prevent reinforcement learning from degrading linguistic fluency or output structure, we insert one SFT refresh round after every five PPO rounds. In total, we conducted 100 rounds of PPO and 20 rounds of SFT training for each offline model.

E.2 Candidate Dialogue Construction and Filtering

The training corpus is built from the diagnostic dialogues described in Appendix A.6. It includes both expert-written seed dialogues and augmented dialogues generated from these seeds. Since not all automatically generated trajectories are equally reliable, we apply reward-based filtering before supervised training.

Specifically, each candidate dialogue is assigned a trajectory-level score based on the same process-oriented criteria later used in reinforcement learning. These criteria include: (1) whether the diagnostic process follows the handbook-defined node progression, (2) whether disease-probability trends are updated in a clinically consistent manner as new evidence emerges, and (3) whether the dialogue contains unsupported claims, including record-inconsistent summaries and guideline-noncompliant reasoning or questioning.

This filtering step serves two purposes. First, it removes low-quality or noisy trajectories before SFT, thereby improving the consistency of the supervision signal. Second, it aligns the demonstration data with the later PPO objective, reducing the mismatch between imitation learning and reinforcement learning.

E.3 Supervised Fine-Tuning initialisation

We use the reward-filtered dialogue set for supervised fine-tuning to initialise the Doctor agent. The objective of this stage is not only to teach the model the correct final diagnosis, but also to establish the desired interaction format and reasoning behaviour. In particular, SFT encourages the model to: (i) follow the structured consultation stages, (ii) ask diagnostically informative follow-up questions, (iii) update disease probabilities in accordance with the handbook and observed evidence, and (iv) produce coherent outpatient-style responses across multiple rounds.

This initialisation is important because pure reinforcement learning from scratch is unstable in

long-horizon dialogue settings. By starting from filtered demonstrations, the Doctor agent already acquires a basic diagnostic policy and a stable linguistic style before online optimisation begins.

E.4 PPO-Based Process optimisation

After SFT initialisation, we optimise the Doctor agent with PPO in a multi-round interactive setting. In each training episode, the Doctor agent interacts with the Patient agent over several consultation rounds. At each round, the Doctor produces the next diagnostic action in natural language, which may include summarising the current case, updating the differential diagnosis, and asking a targeted follow-up question. The Patient agent then responds according to the underlying case record.

Based on the accumulated interaction history, the Teacher agent provides process-level supervision. Its role is to audit whether the current trajectory follows the expected diagnostic process. Concretely, the Teacher checks handbook-node coverage, evaluates whether probability trends of candidate diseases are updated in the correct direction, and identifies unsupported claims. These signals are passed to the Evaluator agent, which transforms them into turn-level rewards and combines them with a final reward based on diagnostic correctness and diagnostic-tree similarity.

Compared with optimising only the final answer, this design allows PPO to improve the consultation trajectory itself. The model is encouraged not only to arrive at the correct diagnosis, but also to do so through a more guideline-consistent, evidence-grounded, and stable reasoning path.

E.5 Interleaved SFT Refresh

To maintain language fluency and structural consistency during PPO training, we periodically perform additional supervised fine-tuning on the filtered dialogue set. Specifically, after every five PPO rounds, we conduct one SFT refresh round.

This design is motivated by a common issue in reinforcement learning for language models: policy optimisation may improve reward-seeking behaviour while gradually harming surface quality, discourse coherence, or output format regularity. In our setting, such degradation is especially undesirable because the Doctor agent is required to maintain a structured consultation format across multiple rounds. Interleaving PPO with SFT helps preserve this structure while still allowing the model to improve process-level decision quality.

E.6 Roles of the Four Agents During Training

The four-agent design separates dialogue generation, simulated response, process auditing, and reward assignment into different functional components.

Doctor Agent. The Doctor agent is the only trainable policy in the framework. It is responsible for conducting the consultation, updating diagnostic hypotheses, and deciding what to ask next.

Patient Agent. The Patient agent simulates the patient's responses based on the underlying structured case record. Its role is to provide consistent and case-grounded answers to the Doctor's questions.

Teacher Agent. The Teacher agent performs process auditing. Given the interaction history, it checks whether the Doctor has covered the expected guideline nodes, whether diagnostic probabilities evolve in a clinically reasonable way, and whether unsupported claims are present.

Evaluator Agent. The Evaluator agent transforms the Teacher's audit outputs and the final consultation outcome into scalar rewards for optimisation. These include both intermediate rewards and final rewards.

E.7 Why Hybrid PPO+SFT Is Needed

A purely supervised approach can teach the model to imitate training trajectories, but it cannot directly optimise the sequential decision quality of the consultation process. Conversely, pure PPO may overfit to sparse rewards and damage the fluency or structural regularity of the generated dialogue. Our hybrid strategy is intended to balance these two objectives.

SFT provides a strong initialisation in terms of language quality, interaction format, and basic diagnostic behaviour. PPO then improves the policy by optimising process-level and outcome-level rewards in online interaction. Periodic SFT refresh acts as a stabiliser that preserves the model's structured output ability. Together, these components enable the Doctor agent to improve both diagnostic performance and process alignment.

E.8 Relation to the Main Experiments

All framework-enhanced Doctor models reported in the main experiments are trained under this hybrid PPO+SFT scheme. In the main paper, we

compare each such model against its corresponding base model under direct prompting. We further conduct ablation studies on three key components of the framework: the companion guideline file, the Teacher agent, and the turn-level optimisation procedure. These ablations isolate the contributions of external process knowledge, structured auditing feedback, and reward-based optimisation, respectively.

F Confusion Matrix for Non-Adherence

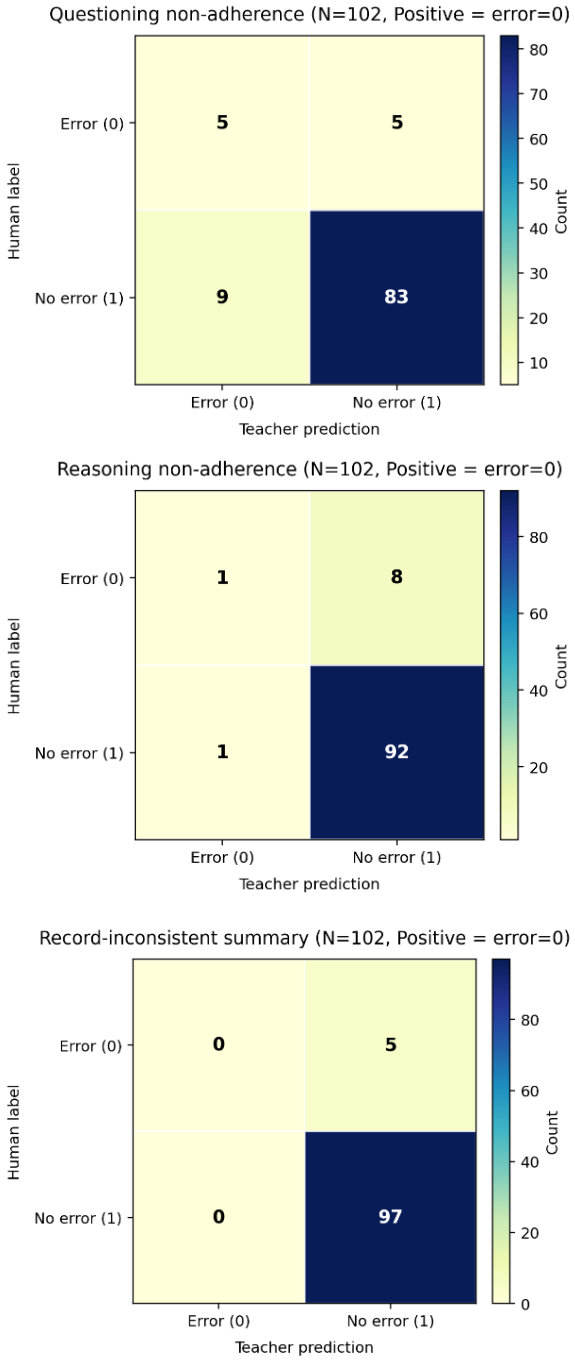


Figure 5: Confusion matrices for questioning non-adherence (top), reasoning non-adherence (middle) and record-inconsistent summary(bottom).

KGRxn-LLM: Knowledge Graph Enhanced Large Language Models for Molecular Reaction Reasoning

Weichen Liu[1] **Qiyao Xue**[2] **Yuyang Wu**[3]
Olexandr Isayev[3] **Natasa Miskov-Zivanov**[1]
[1]University of Pittsburgh, [2]UNC-Chapel Hill, [3]Carnegie Mellon University
weichenliu@pitt.edu, nmzivanov@pitt.edu

Abstract

Large language models (LLMs) demonstrate strong general language capabilities but remain limited in chemical reasoning, particularly for tasks requiring structured, mechanistic understanding of molecular reactions. We present **K**nowledge **G**raph **R**eaction **LLM** (KGRxn-LLM), a framework that augments LLMs with a hierarchical chemical knowledge graph (KG) to ground reasoning in molecular transformations and reaction patterns. Existing benchmarks primarily emphasize reaction or molecular fact recall, providing limited assessment of reaction-level mechanistic reasoning. To address this gap, we introduce **KGRxn-Bench**, a benchmark of 1,200 questions designed to evaluate LLMs on reaction-centric reasoning tasks, including functional group identification, reaction type classification, and product and reagent prediction. Experimental results show that our approach of grounding LLMs in structured KG substantially improves performance across multiple tasks and model backbones, outperforming domain specific fine-tuned models on KG-covered splits and most hold-out splits.[1]

1 Introduction

Recent advances in LLMs have led to remarkable progress on a broad spectrum of common-sense and general purpose language understanding tasks (Touvron et al., 2023; Achiam et al., 2023; Bai et al., 2023). Despite this success, LLMs continue to face significant challenges applied to chemistry, where domain knowledge is highly structured and governed by strict chemical rules rather than simple logic correlations (Bran et al., 2023; Jiménez-Luna et al., 2021). Prior efforts have extended LLMs to chemical domains such as molecular property question answering (QA), molecule caption generation, and molecular description tasks based on SMILES, which is string representation

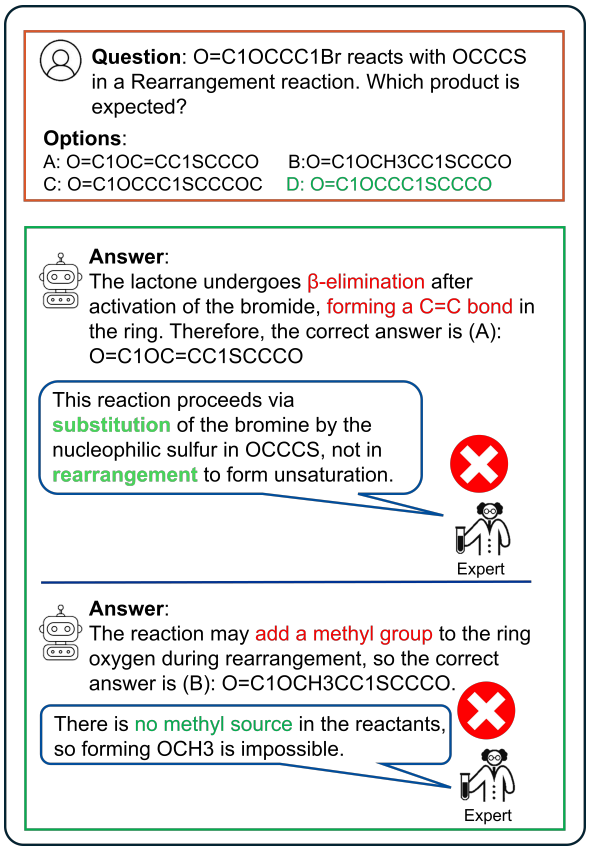


Figure 1: Illustration of hallucination problems of LLM doing molecular reaction reasoning.

of molecular structures using predefined syntax (Lu et al., 2024; Laghuvarapu et al., 2025), demonstrating that LLMs can learn to generate syntactically valid chemical text by fine-tuning on large scale related corpus. However, recent evaluations show that these models remain fundamentally limited in chemical reaction reasoning, often misidentifying functional groups, confusing reaction types, and failing to track structural transformations across reactants and products, especially on tasks that require multi-step reasoning about reaction mechanisms or functional group changes (Runcie et al., 2025; Chen et al., 2025; Zhang et al., 2025).

[1]Code and benchmark are available at https://github.com/pitt-miskov-zivanov-lab/KGRxn-LLM.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 268–281
July 3-4, 2026 ©2026 Association for Computational Linguistics

Current LLMs face hallucination problems in molecular reaction reasoning. Domain-specific models like ChemBERTa, MolT5, and Chemformer (Chithrananda et al., 2020; Edwards et al., 2022; Irwin et al., 2022) benefit from large-scale SMILES pretraining but these models may produce chemically invalid conclusions despite generating explanations that appear internally coherent. As shown in Figure 1, the first answer is confused about the "Rearrangement" reaction mechanism and the second answer involves wrong functional group source in reactant. A knowledge graph helps address this limitation by providing explicit chemical concepts as structured evidence for reasoning. Instead of relying only on pretrained knowledge, the model can retrieve grounded facts about molecules, functional groups, and reactions, which reduces hallucination and makes the reasoning process consistent with chemical rules.

Another fundamental challenge in this domain is the lack of benchmarks that explicitly evaluate LLMs' reasoning ability about molecular reaction transformations. Existing USPTO-based benchmarks (Chen et al., 2024; Yuan et al., 2025) focus primarily on forward reaction prediction or retrosynthesis, where success is measured by whether a model generates the correct product or SMILES representation. While these tasks are valuable, they do not assess whether a model understands the underlying transformation mechanisms or how functional groups evolve between reactants and products. On the other hand, chemical QA datasets such as MoleculeQA (Lu et al., 2024) and MolTextQA (Chen et al., 2025) focus on factual property comprehension, leaving reaction level reasoning ability such as identifying reaction types, tracking structural changes, or explaining mechanism relevant functional group operations largely untested.

To address these gaps, we introduce KGRxn-LLM, a knowledge graph(KG) integrated framework designed to advance knowledge-grounded molecular reasoning with LLMs. Our contribution can be summarized as follows:

- KGRxn-LLM introduces a hierarchical chemical knowledge graph with multiple layers, containing molecules, functional groups, reaction types and reaction instances. This KG encodes both chemical relationships and structural attributes of the chemical concept, forming an interpretable knowledge substrate.
- We propose KGRxn-Bench, a molecular reasoning benchmark built from curated USPTO reactions and annotated with functional group mappings and reactant-to-product transformations. The benchmark comprises tasks that evaluate reaction type identification, functional group transformation reasoning, and product prediction.
- We also propose an agent-based retrieval workflow that efficiently extracts transformation-relevant knowledge such as functional groups, common reaction patterns, similar molecules and reaction analogues, and uses this information to ground LLM reasoning, reducing hallucinations and improving accuracy.

2 Related work

LLMs for Biochemical Reasoning. The application of LLMs in biochemistry and computational biology has progressed rapidly, shifting from general text processing to specialized scientific discovery and molecular reasoning (Han et al., 2025; Usuyama et al., 2025). Recent paradigms treat biological molecules and small organic compounds as distinct text structures, allowing LLMs to excel in tasks ranging from protein structure prediction and *de novo* protein design to complex genomic analyses. For example, recent agent-based workflows have integrated LLMs to generate interpretable hypotheses for optimizing protein binding motifs and biochemical structures (Akke et al., 2025). Furthermore, the deployment of specialized multi-agent systems has enabled the autonomous execution of complex biochemical tasks, such as automating molecular dynamics simulations through tools (Campbell et al., 2025).

Parallel to these agentic frameworks, the emergence of advanced reinforcement learning-based reasoning models, such as DeepSeek-R1 and GPT-o series models (Guo et al., 2025; Jaech et al., 2024) has pushed the boundaries of zero-shot chemical intelligence, demonstrating robust capabilities in structural elucidation and the direct interpretation of molecular trends (Cui et al., 2025). While these models demonstrate impressive proficiency in broad biochemical property prediction and hypothesis generation, they continue to struggle with the rigorous, step-by-step mechanistic reasoning required to track atomic-level functional group transformations during chemical reactions. This fundamental limitation underscores the necessity for structurally knowledge-grounded approaches.

Knowledge Graph Guided LLM Reasoning. To mitigate hallucination and improve the factual reliability of LLMs, knowledge graph integration for Retrieval-Augmented Generation (RAG) has been employed (Lewis et al., 2020; Pan et al., 2024). Broadly, LLMs incorporate KGs to enhance reasoning through three primary mechanisms, including representation fusion, prompt augmentation, and tool-augmented querying (Pan et al., 2024). In representation fusion, KG network embeddings are injected directly into the LLMs' latent space during pre-training or fine-tuning, explicitly aligning textual tokens with structural entities to improve the model's internal knowledge representations (Yasunaga et al., 2022). Conversely, prompt-based retrieval methods, such as GraphRAG, dynamically extract multi-hop subgraphs, community summaries, or entity-relation triples from extensive knowledge bases, appending this structured context to the LLMs' input to guide explicit, step-by-step reasoning (Edge et al., 2024). To handle more complex logical queries, recent agentic frameworks like Think-on-Graph (ToG) and Reasoning-on-Graphs (RoG) empower LLMs to act as autonomous agents that iteratively explore graph structures by generating executable SQL queries to traverse multi-hop relational pathways before synthesizing an answer (Sun et al., 2023; Luo et al., 2023).

Within biochemical domains, these general frameworks have been successfully adapted by linking LLMs with specialized KGs such as UMLS or DrugBank to advance relational tasks like biomedical question answering and drug-target interaction prediction (Jin et al., 2024). Even though these domain-specific graphs excel at retrieving facts, they predominantly model knowledge as static nodes. Consequently, they remain unsuitable for capturing the dynamic, atomic-level transformations and mechanistic pathways required for complex chemical reaction reasoning. Our framework solves this problem with agentic retrieving algorithm, leveraging KG with comprehensive relationships of molecules and reactions.

Molecular Reasoning Benchmarks. The rapid integration of LLMs into biochemistry has motivated the development of specialized benchmarks for systematically evaluating their domain knowledge and reasoning capabilities. Recent benchmarks such as ChemLLMBench and ChemIQ (Guo et al., 2023; Runcie et al., 2025) assess a broad range of general chemistry tasks, including molecular property prediction, IUPAC to SMILES translation, and factual knowledge recall. To evaluate deeper structural understanding beyond conventional question answering, more recent benchmarks have begun to focus on fine grained and verifiable structural reasoning tasks. For example, FGBench introduces a rigorous evaluation framework for property reasoning at the functional group level, testing whether an LLM can interpret both single group and multi group structural interactions (Liu et al., 2025). ChemCoTBench tests stepwise molecular editing and optimization (Li et al., 2025), and MolErr2Fix measures the ability to detect and explain semantic hallucinations in molecular descriptions (Wu et al., 2025). However, these benchmarks provide limited coverage of reaction-centric reasoning, particularly mechanistic understanding of reactant-to-product transformations.

3 KGRxn-LLM Framework

3.1 Data Processing and Annotation

Preliminary Processing In retrieval-augmented settings, an excessively large graph may introduce noisy evidence, reduce retrieval precision, and degrade downstream reasoning quality (Gao et al., 2023). Therefore, we balance graph coverage with retrieval reliability. Following prior reaction prediction work (Coley et al., 2019), we extract 32,000 reactions from the USPTO reaction dataset (Marco et al., 2015). Among them, 30,000 reactions are used to construct the knowledge graph and the KG-covered portion of KGRxn-Bench, while the remaining 2,000 reactions are held out for constructing the benchmark hold-out split.

Starting from raw USPTO records, we first convert each reaction into a structured instance through JSON parsing and schema normalization, extracting reactants, products, and reaction-level metadata into a unified format. As illustrated in Figure 2, we then use RDKit to canonicalize molecular structures, trace atom correspondences between reactants and products, and identify functional groups in each molecule. Based on the atom-mapping results, we characterize functional-group-level transformations by determining which groups are retained, removed, or introduced during the reaction. The processed reaction, molecule, and functional-group annotations are stored in a structured database, which serves as the basis for knowledge graph construction and the KGRxn-LLM reasoning framework. More details are provided in Appendix A.1.

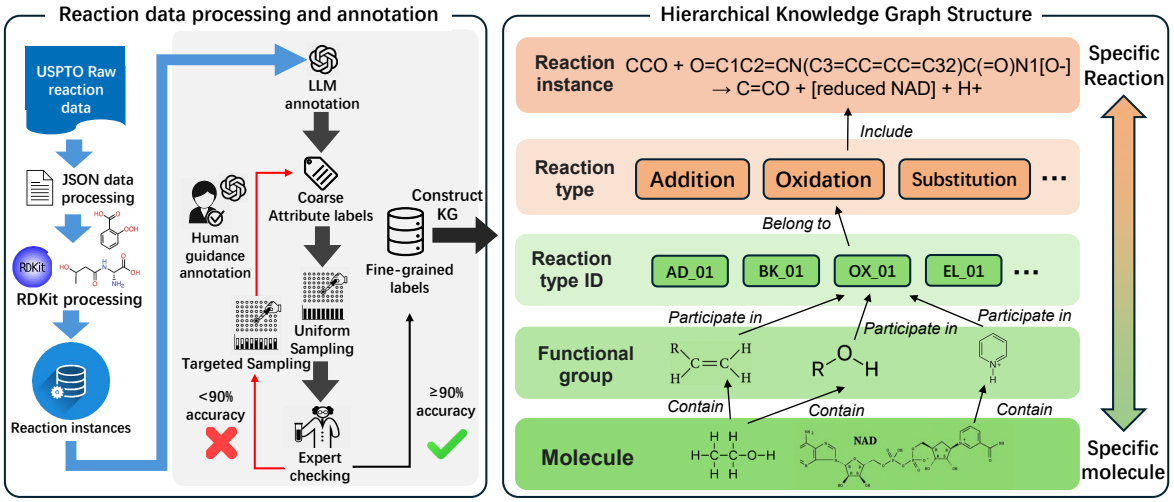


Figure 2: Overview of KGRxn-LLM data annotation pipeline and KG structure. Left: Raw USPTO reactions are processed into structured reaction instances, then annotated through an LLM-assisted pipeline with expert checking and resampling until reaching satisfactory accuracy. Right: Our hierarchical knowledge graph links molecules, functional groups, reaction type IDs, reaction types, and specific reaction instances across multiple levels.

Iterative Refinement Annotation Pipeline As shown in Figure 2, we employ an iterative human-in-the-loop pipeline to annotate KG entities and associated semantic labels in a scalable yet quality-controlled manner. In the first round, a strong closed-source LLM (ChatGPT 5) is used to produce coarse annotations for graph elements involved in knowledge graph construction, including reaction types, functional group categories and properties, and molecule-level properties. To assess annotation reliability, we uniformly sample a subset of instances for expert review and estimate annotation quality by empirical accuracy:

$$A = \frac{1}{|\mathcal{S}|} \sum_{x_i \in \mathcal{S}} \mathbf{1}[\hat{y}_i = y_i], \tag{1}$$

where $\mathcal{S}$ denotes the expert-reviewed subset, $\hat{y}_i$ is the LLM annotation, and y_i is the expert label.

When the verified accuracy falls below a predefined threshold of 90%, we perform a new round of targeted sampling that prioritizes instances from categories with higher observed error rates, while preserving the class composition of the previously reviewed subset to avoid severe sampling drift. Specifically, for each category c, we assign a sampling weight: $w_c \propto \pi_c(1 + \lambda e_c)$, where π_c is the category proportion in the previous review subset, e_c is the corresponding expert-verified error rate, and λ controls the emphasis on error-prone categories. This strategy increases the chance of selecting rare reaction types, molecules with multiple plausible functional groups, and structurally similar labels that are easy to confuse, while preserving the overall class composition.

At each audit round, previous observed errors are used to refine the annotation protocol with corrected examples, category-specific constraints, and explicit decision rules for ambiguous cases. The affected instances are then re-annotated under the revised protocol and audited again. This audit–refinement cycle is repeated until the target accuracy is reached. We also record category-wise error rates to identify systematic failure modes, allowing the KG annotation pipeline to scale beyond fully manual labeling while reducing bias from the pre-annotation model.

3.2 Reaction Knowledge Graph Construction

We construct the reaction knowledge graph as a hierarchical structure that organizes chemical knowledge from concrete molecular structures to abstract reaction semantics. Instead of treating each reaction as an isolated reactant–product pair, the graph decomposes reaction knowledge into multiple interconnected layers, including molecules, functional groups, fine-grained reaction patterns, coarse reaction categories, and specific reaction instances. This design improves retrieval efficiency by explicitly linking structural evidence to higher-level reaction concepts.

At the lowest level, molecule nodes represent the specific compounds involved in reactions. These

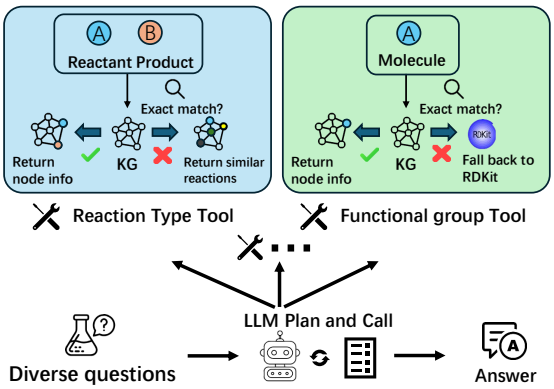


Figure 3: Task-specific tool calling and two-stage retrieval in KGRxn-LLM.

nodes preserve the original structural information and serve as the foundation of the graph. On top of this, functional group nodes capture chemically meaningful substructures, such as alcohols, alkenes, or carbonyl groups, and are linked to molecule nodes through "contain" relations. This layer provides a more reaction-relevant abstraction, since many reaction outcomes are determined by functional group transformations rather than whole-molecule identity.

Above the functional group layer, we introduce fine-grained Reaction Type ID nodes, which represent specific transformation patterns defined by characteristic local structural or functional-group changes. These nodes are further grouped into coarse reaction type nodes, such as oxidation, addition, and substitution, to provide a higher-level semantic taxonomy of reactions. Finally, reaction instance nodes represent individual reactions in the dataset and connect these abstract categories back to concrete examples. Overall, this hierarchical organization enables the graph to support both bottom-up reasoning from molecular structure to reaction type and top-down retrieval from reaction classes to plausible molecular transformations. More details can be found at Appendix A.2.

3.3 Knowledge Graph Retrieval Method

As illustrated in Figure 3, we formulate retrieval as an agentic tool-use process rather than a fixed lookup pipeline. Given a molecular reaction question, the LLM is prompted to determines which operation is required and then invokes the corresponding tool, such as functional group identification, reaction type inference, or other reaction-oriented utilities. Each tool is connected to the knowledge graph and follows a common two-stage retrieval strategy. It first attempts exact retrieval by matching canonicalized molecular or reaction representations to graph nodes, in which case the associated graph facts are returned directly as high-confidence evidence. If no exact match is available, the tool falls back to analogy-based retrieval at the functional-group level. Specifically, the query is represented by its consumed and formed functional group sets, and candidate reactions are ranked by the average Jaccard similarity (Travieso et al., 2024)

$$J_{\text{sim}} = \frac{1}{2}\left(\frac{|C_q \cap C_r|}{|C_q \cup C_r|} + \frac{|F_q \cap F_r|}{|F_q \cup F_r|}\right), \quad (2)$$

where C_q and F_q denote the consumed and formed functional groups of the query, and C_r and F_r denote those of a retrieved reaction. This retrieval criterion emphasizes analogous transformation patterns rather than superficial molecular similarity. For molecule-level queries such as functional group identification, the fallback is performed directly with RDKit-based functional group matching. The retrieved evidence is prepended to the original question and passed to the LLM for final answer generation in a single inference step. This design allows the model to adaptively choose the most relevant tool for each question while grounding its reasoning in explicit chemical evidence from either exact graph matches or structurally similar reaction analogues.

3.4 KGRxn-Bench Construction

We construct *KGRxn-Bench*, a benchmark designed to evaluate LLMs' ability on molecular reaction reasoning. As summarized in Table 1, the benchmark contains 1,200 questions derived from 986 unique reactions, covering 3,850 unique molecules, 10 reaction types, and 160 functional groups. The benchmark is designed to probe multiple levels of chemical reasoning, ranging from functional group recognition to reaction-level inference and condition prediction.

Task Definitions. KGRxn-Bench consists of four task types, each targeting a distinct aspect of reaction understanding.

Functional group identification requires the model to determine which functional group is present in a molecule represented by a SMILES string. This task evaluates whether the model can recognize chemically meaningful substructures from molecular representations.

Statistics	Number
Total Questions	1200
Unique Reactions	986
Unique Molecules	3850
Unique Reaction Types	10
Unique Functional Groups	160
Question Type Distribution	
Functional Group Identification (KG-covered)	272 (22.7%)
Functional Group Identification (hold-out)	100 (8.3%)
Reaction Type Classification (KG-covered)	243 (20.3%)
Reaction Type Classification (hold-out)	100 (8.3%)
Reaction Product Prediction (KG-covered)	151 (12.6%)
Reaction Product Prediction (hold-out)	100 (8.3%)
Reagent Prediction (KG-covered)	134 (11.2%)
Reagent Prediction (hold-out)	100 (8.3%)

Table 1: Key statistics of KGRxn-Bench.

Reaction type classification requires the model to infer the reaction category from reactant and product SMILES. Given the structural transformation, the model must assign the reaction to a predefined class such as substitution, oxidation, or addition. This task evaluates whether the model can abstract reaction semantics from explicit molecular changes.

Reaction product prediction requires the model to predict the expected product given two reactants and a specified reaction type. Unlike reaction classification, this task requires forward reasoning about how functional groups are transformed under a given reaction class.

Reagent prediction requires the model to infer plausible reagents (E.g.catalysts, solvents) from the observed reactant-to-product transformation. This task tests whether the model can associate structural changes with the reaction conditions typically required to realize them.

Question Answer Pair Generation. Questions are generated programmatically from the processed reaction data described in Section 3.1. For functional group identification, reaction type classification, and reagent prediction, multiple-choice distractors are generated by randomly sampling from the set of valid labels observed in the dataset. Specifically, we sample three incorrect options from the corresponding global label space, excluding the ground-truth answer, and combine them with the correct answer to form a multiple choice question with 4 options.

For reaction product prediction, we instead adopt a structure-aware mutation strategy to construct chemically plausible distractors. Rather than drawing unrelated molecules, we generate negative options by applying targeted modifications to the ground-truth product SMILES. These modifications include halogen substitution, replacement of carbonyl-adjacent groups, functional group swapping, aromatic position isomerization, double-bond toggling, substituent deletion, and small substituent addition. Mutation operators are applied iteratively until three unique and chemically valid distractors are obtained. This design makes the task substantially more challenging, as the incorrect options remain structurally similar to the correct answer and therefore require genuine mechanistic reasoning to distinguish.

KG-covered and Hold-out Splits. To evaluate both retrieval effectiveness and out-of-graph generalization, we partition the benchmark into *KG-covered* and *hold-out* subsets. The KG-covered subset contains questions derived from reactions included in the knowledge graph, allowing the retrieval module to access exact or closely related reaction evidence. In contrast, the hold-out subset is constructed from 2,000 reactions excluded during KG construction, ensuring that no exact reaction match is available in the graph.

Each task contains 100 hold-out questions, while the number of KG-covered questions varies across tasks according to the availability of eligible reactions. These two splits allow KGRxn-Bench to evaluate not only whether a model can benefit from graph-grounded retrieval, but also whether it can generalize to unseen reactions through higher level chemical similarity.

4 Experiments

4.1 Experiment Settings

Benchmarks We evaluate KGRxn-LLM on our proposed KGRxn-Bench together with two external benchmarks, FGBench (Liu et al., 2025) and SciKnowEval (Feng et al., 2024). KGRxn-Bench serves as our primary benchmark for molecular reaction reasoning, containing 1,200 multiple-choice questions spanning four tasks. We report results on both the KG-covered split and the hold-out split to evaluate the framework's ability to benefit from graph-grounded retrieval as well as to generalize beyond directly covered reactions. To further assess generalization beyond reaction-centered tasks, we additionally evaluate on FGBench and SciKnowEval, which probe broader biochemical understanding. Specifically, on FGBench we consider the Boolean tasks of single functional group impact,

	Functional Group		Reaction Product		Reaction Type		Reagent Prediction	
	KC	HO	KC	HO	KC	HO	KC	HO
# of samples	272	100	151	100	243	100	134	100
Vanilla Models								
Llama-3.1-8B	58.82	60.00	48.34	43.00	48.56	56.00	8.21	19.00
Qwen2.5-7B	63.97	67.00	48.34	63.00	46.09	64.00	14.18	13.00
nach0-base	19.12	20.00	18.54	12.00	6.17	6.00	9.70	7.00
LlaSMol-Mistral-7B	34.93	35.00	28.48	29.00	21.81	34.00	28.36	27.00
ChemLLM-7B	67.65	69.00	41.06	38.00	37.45	44.00	26.12	27.00
ChemLLM-7B-DPO	82.46	77.00	**73.51**	**71.00**	42.67	51.00	53.73	**40.00**
ChemDFM-v1.5-8B	67.65	77.00	43.71	46.00	61.32	<u>65.00</u>	25.37	33.00
KGRxn-LLM (Ours)								
Llama-3.1-8B + KG	<u>94.12</u>	**94.00**	<u>72.85</u>	56.00	**95.88**	64.00	**96.27**	37.00
Qwen2.5-7B + KG	**97.79**	<u>92.00</u>	**73.51**	<u>70.00</u>	<u>90.95</u>	**69.00**	<u>95.52</u>	<u>39.00</u>

Table 2: Performance in accuracy on KGRxn-Bench across four chemical reaction reasoning tasks. KC = KG-covered set, HO = Hold-out set. Best model results are in **bold**. Second best model results are <u>underlined</u>.

multiple functional group interactions and molecular comparisons. On SciKnowEval, we evaluate both multiple-choice and boolean questions in chemistry and biology. We use zero-shot accuracy as the evaluation metric for all tasks.

Baselines We compare KGRxn-LLM against both general-purpose and domain-specific language models. Specifically, we include instruction-tuned LLMs such as Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct as general model baselines, as well as chemistry-specialized models including nach0-base, LlaSMol-Mistral-7B, ChemLLM-7B, ChemLLM-7B-DPO and ChemDFM-v1.5-8B (Livne et al., 2024; Yu et al., 2024; Zhang et al., 2024; Zhao et al., 2025). All models are evaluated under a consistent zero-shot prompting setting without additional fine-tuning.

Implementation Details The knowledge graph is constructed from 30,000 USPTO reactions and hosted on a Neo4j Aura cloud instance. The graph contains overall 146,500 nodes and 269,981 relationships linking through the five-level graph hierarchy. For the retrieval module, the Jaccard similarity threshold is set to 0.3, with a maximum of 5 retrieved analogy reactions to ensure enough information guidance from KG. During generation, we use a temperature of 0.2, top-p of 0.9, and a maximum of 1024 new tokens for all question to accommodate long reasoning chains. All models are prompted with a zero-shot chain-of-thought format. More detailed prompt can be found in Appendix B.

4.2 Experimental results

As shown in Table 2, KGRxn-LLM consistently outperforms most baseline models across tasks on KGRxn-Bench. Performance gains are particularly substantial on the KG-covered split, where knowledge retrieval yields massive improvements over vanilla models. For instance, KG augmentation increases Llama-3.1-8B accuracy on functional group identification from 58.82% to 94.12%.This result confirms that when high-quality structured evidence is available, even general-purpose LLMs can leverage it effectively. Similar trends across other tasks confirm that structured knowledge effectively enhances mechanistic reasoning. On the hold-out split, KGRxn-LLM maintains consistent gains, demonstrating its capacity to generalize. Specifically, the augmented Qwen2.5-7B model improves functional group identification from 67.00% to 92.00%, suggesting that retrieved structural evidence provides a valuable reference even without exact graph matches.

Furthermore, Table 3 shows that KGRxn-LLM also improves performance on the external benchmarks FGBench and SciKnowEval. On FGBench, augmentation elevates functional group reasoning, notably raising the Qwen2.5-7B score from 39.33% to 49.64% on interaction queries. On SciKnowEval, the augmented Llama model improves from 88.24% to 95.76% on the biological boolean subset, reflecting broad gains across both chemistry and biology. Collectively, these findings indicate that the proposed framework benefits both specialized reaction

Model	FGBench			SciKnowEval			
	Single	Inter.	Comp.	Chem MCQ	Chem boolean	Bio MCQ	Bio boolean
Llama-3.1-8B	68.36	43.67	45.52	63.41	**78.67**	66.61	88.24
Qwen2.5-7B	57.34	39.33	**51.64**	68.28	66.67	76.23	83.36
Llama-3.1-8B + KG	**70.83**	43.95	46.94	71.71	77.25	68.35	**95.76**
Qwen2.5-7B +KG	68.50	**49.64**	48.32	**72.97**	76.53	**77.19**	92.79

Table 3: Performance in accuracy on external benchmarks: FGBench and SciKnowEval. In FGBench we consider the Boolean tasks of single functional group impact, multiple functional group interactions and molecular comparisons. In SciKnowEval we evaluate Multiple Choice Question(MCQ) and Boolean on Chemistry and biology subset. Best results per column are in **bold**.

reasoning and broader biochemical understanding through transferable structured knowledge.

4.3 Analysis

Overcoming Structural Hallucination A primary challenge for standard LLMs in chemistry is the tendency to generate structural hallucinations, where the model suggests products containing atoms or functional groups not present in the initial reactants. As illustrated in the comparative analysis, vanilla models often fail to maintain a strict atom inventory, erroneously adding substituents like methyl groups during rearrangement tasks despite no such source existing in the input SMILES. KGRxn-LLM mitigates this by providing pre-computed functional group annotations and structural transformations from KG. By grounding the reasoning process from different molecules to specific functional groups, the framework ensures that the model's internal logic remains consistent with chemical constraints.

Generalization Through Analogy The framework demonstrates a capacity of chemical reasoning by analogy, particularly evident in its performance on hold-out reaction sets where exact matches are unavailable in the KG. When a direct lookup fails, the agentic tool-use process falls back to a two-stage retrieval strategy that calculates Jaccard similarity based on consumed and formed functional group sets. This approach allows the LLM to identify analogous transformation patterns from similar reactions, such as the conversion of an alcohol to an ester, and apply those underlying rules to unseen molecular structures. Experimental results on the hold-out split confirm this effectiveness, as the retrieval of structurally similar reaction analogues provides a reliable reference that guides the model toward correct mechanistic pathways even in extrapolative settings.

Beyond Reaction Centered Tasks Result on FGBench reveals that vanilla models struggle with multi functional group interactions, while knowledge augmentation grounds these relationships in verified chemical rules. Similarly on SciKnowEval, the graph supplies foundational context that prevents factual error in broader chemistry and biology queries. Although it occasionally happens that in general knowledge evaluation settings performance drop compared to vanilla model when KG fail to retrieve any irrelevant context, this problem can be solved by include broader range of knowledge.

5 Conclusion

In this work, we present KGRxn-LLM, a knowledge graph augmented framework for molecular reaction reasoning that integrates hierarchical chemical knowledge with functional group level information retrieval. We also introduce KGRxn-Bench, a benchmark designed to evaluate reaction-centric reasoning beyond standard knowledge recalling tasks. Experimental results demonstrate that our framework of augmenting LLMs with structured chemical knowledge graph substantially improves performance across multiple reasoning tasks and model backbones, while also generalizing to broader biochemical understanding.

Limitations

This study has several limitations. First, the adopted taxonomy of ten coarse-grained reaction types provides only a simplified abstraction of organic chemistry. Second, the current knowledge graph, constructed from approximately 30K reactions, remains modest in scale relative to the vast space of known chemical reactions, which may constrain both coverage and retrieval effectiveness. Finally,

the framework has only been evaluated on mid-scale language models (7–8B parameters), and its scalability and effectiveness on larger models remain to be systematically investigated.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mattias Akke, Soojung Yang, Jurgis Ruza, Jinyeop Song, Elton Pan, and Rafael Gomez-Bombarelli. 2025. Bayesian optimization for biochemical discovery with llms.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Quintina Campbell, Sam Cox, Jorge Medina, Brittany Watterson, and Andrew White. 2025. Mdcrow: Automating molecular dynamics workflows with large language models. *Machine Learning: Science and Technology*.

Shuan Chen, Ramil Babazade, Taewan Kim, Sunkyu Han, and Yousung Jung. 2024. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Scientific Data*, 11(1):863.

Yufan Chen, Ching Ting Leung, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. 2025. Towards large-scale chemical reaction image parsing via a multimodal large language model. *arXiv preprint arXiv:2503.08156*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377.

Dong-Xu Cui, Shi-Yu Long, Yi-Xuan Tang, Yue Zhao, and Qiao Li. 2025. Can reasoning power significantly improve the knowledge of large language models for chemistry? based on conversations with llms. *Journal of Chemical Information and Modeling*, 65(18):9516–9527.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in neural information processing systems*, 36:59662–59688.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. 2025. From generalist to specialist: A survey of large language models for chemistry. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1106–1123.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

José Jiménez-Luna, Francesca Grisoni, Nils Weskamp, and Gisbert Schneider. 2021. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert opinion on drug discovery*, 16(9):949–959.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

Siddhartha Laghuvarapu, Namkyeong Lee, Chufan Gao, and Jimeng Sun. 2025. Moltextqa: A question-answering dataset and benchmark for evaluating multimodal architectures and llms on molecular structure–text understanding. *Journal of Data-centric Machine Learning Research*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. 2025. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *arXiv preprint arXiv:2505.21318*.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. 2025. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *arXiv preprint arXiv:2508.01055*.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, and 1 others. 2024. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. *arXiv preprint arXiv:2403.08192*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Alan C Marco, Amanda Myers, Stuart JH Graham, Paul D'Agostino, and Kirsten Apple. 2015. The uspto patent assignment dataset: Descriptions and analysis.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. 2025. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 66(1):216–227.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gonzalo Travieso, Alexandre Benatti, and Luciano da F Costa. 2024. An analytical approach to the jaccard similarity index. *arXiv preprint arXiv:2410.16436*.

Naoto Usuyama, Cliff Wong, Sheng Zhang, Tristan Naumann, and Hoifung Poon. 2025. Biomedical natural language processing in the era of large language models. *Annual Review of Biomedical Data Science*, 8.

Yuyang Wu, Jinhui Ye, Shuhao Zhang, Lu Dai, Yonatan Bisk, and Olexandr Isayev. 2025. Molerr2fix: Benchmarking llm trustworthiness in chemistry via modular error detection, localization, explanation, and correction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19365–19382.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Shen Yuan, Shukai Gong, and Hongteng Xu. 2025. Uspto-llm: A large language model-assisted information-enriched chemical reaction dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 817–820.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and 1 others. 2025. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38.

Zihan Zhao, Bo Chen, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, and 1 others. 2025. Chemdfm-r: An chemical reasoner llm enhanced with atomized chemical knowledge. *arXiv e-prints*, pages arXiv–2507.

A Appendix

A.1 Preliminary Processing

Starting from raw USPTO reaction records in JSON format, we extract reaction identifiers, DOI provenance, atom-mapped reaction SMILES, and all input components (reactants, solvents, catalysts) and products into a unified tabular representation. All molecular SMILES are canonicalized using RDKit, and each molecule is assigned a deterministic identifier via the SHA-1 hash of its canonical SMILES (first 16 hex characters), ensuring consistent deduplication across reactions. We then apply 220 SMARTS-based substructure queries to detect functional groups in each molecule. When multiple patterns match the same atom set, a specificity ranking retains only the most descriptive label (e.g., *alcohol_secondary* over generic *alcohol*). For atom-mapped reactions, we identify the reaction center by comparing atom-level properties between reactant and product sides, then expand the changed-atom set by one bond radius. Only functional group instances overlapping this expanded center are regarded as reaction-participating groups. We then compute *formed* and *consumed* functional group sets by comparing (name, atom-map tuple) pairs across sides, which serve as input to a rule-based reaction type classifier containing 34 prioritized rules across seven priority levels that assign each reaction to one of ten coarse types. Reactions matching no rule are labeled *Unknown* and excluded. Finally, molecules originally annotated as reactants but who do not participate in the reaction center are reclassified as unchanged agents.

```
FG_SMARTS = {
    "alkene": Chem.MolFromSmarts("C=C"),
    "alkyne": Chem.MolFromSmarts("C#C"),
    "allene": Chem.MolFromSmarts("C=C=C"),
    "cumulene": Chem.MolFromSmarts("C(=C)=C"),
    "carbocation": Chem.MolFromSmarts("[C+]"),
    "carbanion": Chem.MolFromSmarts("[C-]"),
    "carbene": Chem.MolFromSmarts("[CX2]"),
    "borane": Chem.MolFromSmarts("[BH3]"),
    "boronic_acid": Chem.MolFromSmarts("B(O)O"),
    "boronic_ester": Chem.MolFromSmarts("B(O[#6])O[#6]"),
    "borinic_acid": Chem.MolFromSmarts("B(O)[#6]"),
    "borinic_ester": Chem.MolFromSmarts("B(O[#6])[#6]"),
    "borate_ester": Chem.MolFromSmarts("B(O[#6])(O[#6])O[#6]"),
    "alcohol": Chem.MolFromSmarts("[OX2H][#6]"),
    "alcohol_primary": Chem.MolFromSmarts("[CX4H2][OX2H]"),
    "alcohol_secondary": Chem.MolFromSmarts("[CX4H1]([#6])[OX2H]"),
    "alcohol_tertiary": Chem.MolFromSmarts("[CX4]([#6])([#6])[OX2H]"
      ),
    "alkoxide": Chem.MolFromSmarts("[O-][#6]"),
    "ether": Chem.MolFromSmarts("[OD2]([#6])[#6]"),
    "aryl_ether": Chem.MolFromSmarts("[OD2]([#6])[c]"),
    "phenol": Chem.MolFromSmarts("[c][OX2H]"),
    "phenolate": Chem.MolFromSmarts("[c][O-]"),
    "enol": Chem.MolFromSmarts("[CX3](=C)[OX2H]"),
    "enolate": Chem.MolFromSmarts("[CX3](=C)[O-]"),
    "enol_ether": Chem.MolFromSmarts("[CX3](=C)O[#6]"),
    "alkynol": Chem.MolFromSmarts("[CX2]#C[OX2H]"),
    "alkynolate": Chem.MolFromSmarts("[CX2]#C[O-]"),
    "alkynol_ether": Chem.MolFromSmarts("[CX2]#CO[#6]"),
    "ketone": Chem.MolFromSmarts("[CX3](=O)[#6]"),
    "aldehyde": Chem.MolFromSmarts("[CX3H1](=O)[#6]"),
    "ketene": Chem.MolFromSmarts("C=C=O"),
    "hemiketal": Chem.MolFromSmarts("[CX4]([OX2H])([OD2])[#6]"),
    "hemiacetal": Chem.MolFromSmarts("[CX4H1]([OX2H])([OD2])[#6]"),
    "ketal": Chem.MolFromSmarts("[CX4]([OD2])([OD2])[#6]"),
    "acetal": Chem.MolFromSmarts("[CX4H1]([OD2])([OD2])[#6]"),
    "carboxylic_acid": Chem.MolFromSmarts("[CX3](=O)[OX2H1]"),
    "carboxylate": Chem.MolFromSmarts("[CX3](=O)[O-]"),
    "ester": Chem.MolFromSmarts("[CX3](=O)O[#6]"),
    "organic_anhydride": Chem.MolFromSmarts("[CX3](=O)O[CX3](=O)"),
    "carboxylic_anhydride": Chem.MolFromSmarts("[CX3](=O)O[CX3](=O)"
      ),
    "organic_carbonate": Chem.MolFromSmarts("[CX3](=O)(O[#6])O[#6]"),

    "organic_hydroperoxide": Chem.MolFromSmarts("[OX2H]O[OX2]"),
    "organic_peroxide": Chem.MolFromSmarts("[OX2]O[OX2]"),
    "peroxyacid": Chem.MolFromSmarts("[CX3](=O)OO[OX2H]"),
    "methylenedioxy": Chem.MolFromSmarts("OCOc"),
    "ethylenedioxy": Chem.MolFromSmarts("OCCOc"),
    "epoxy": Chem.MolFromSmarts("C1OC1"),
    "silane": Chem.MolFromSmarts("[SiH4]"),
    "siloxane": Chem.MolFromSmarts("[Si]O[Si]"),
    "silyl_ether": Chem.MolFromSmarts("[Si]O[#6]"),
    "silyl_enol_ether": Chem.MolFromSmarts("[Si]O[C]=C"),
    "silyl_alkynol_ether": Chem.MolFromSmarts("[Si]O[C]#C"),
    "phosphine": Chem.MolFromSmarts("P([#6])([#6])[#6]"),
    "phosphonium": Chem.MolFromSmarts("[P+]"),
    "aminophosphine": Chem.MolFromSmarts("P(N)([#6])[#6]"),
    "phosphine_oxide": Chem.MolFromSmarts("P(=O)([#6])([#6])[#6]"),
    "phosphinic_acid": Chem.MolFromSmarts("P(=O)(O)[#6]"),
    "phosphinate": Chem.MolFromSmarts("P(=O)([O-])[#6]"),
    "phosphonic_acid": Chem.MolFromSmarts("P(=O)(O)(O)[#6]"),
    "phosphonate": Chem.MolFromSmarts("P(=O)(O[#6])([O-])[#6]"),
    "phosphite_ester": Chem.MolFromSmarts("P(O[#6])(O[#6])O[#6]"),
    "phosphinite": Chem.MolFromSmarts("P(O[#6])([#6])[#6]"),
    "phosphonite": Chem.MolFromSmarts("P(O[#6])(O[#6])[#6]"),
    "phosphodiester": Chem.MolFromSmarts("P(=O)(O[#6])O[#6]"),
    "phosphate_mono_ester": Chem.MolFromSmarts("P(=O)(O)(O)O[#6])"),
    "phosphate_tri_ester": Chem.MolFromSmarts("P(=O)(O[#6])(O[#6])O
      [#6]"),
    "phosphoramide": Chem.MolFromSmarts("P(=O)(N)(O[#6])O[#6]"),
    "thiophosphate": Chem.MolFromSmarts("P(=S)(O[#6])(O[#6])O[#6]"),
    "phosphonium_ylide": Chem.MolFromSmarts("[P+][C-]"),
    "fluoro": Chem.MolFromSmarts("[F]"),
    "chloro": Chem.MolFromSmarts("[Cl]"),
    "bromo": Chem.MolFromSmarts("[Br]"),
    "iodo": Chem.MolFromSmarts("[I]"),
    "alkyl_halide": Chem.MolFromSmarts("[CX4][F,Cl,Br,I]"),
    "vinyl_halide": Chem.MolFromSmarts("[CX3]=[CX3][F,Cl,Br,I]"),
    "aryl_halide": Chem.MolFromSmarts("[c][F,Cl,Br,I]"),
    "halamine": Chem.MolFromSmarts("N[Cl,Br,I,F]"),
    "sulfenyl_halide": Chem.MolFromSmarts("S[Cl,Br,I,F]"),
    "sulfinyl_halide": Chem.MolFromSmarts("S(=O)[Cl,Br,I,F]"),
    "sulfonyl_halide": Chem.MolFromSmarts("S(=O)(=O)[Cl,Br,I,F]"),
    "halosulfate": Chem.MolFromSmarts("OS(=O)(=O)[Cl,Br,I,F]"),
    "phosphoryl_halide": Chem.MolFromSmarts("P(=O)([Cl,Br,I,F])[Cl,
      Br,I,F]"),
    "phosphorus_halide": Chem.MolFromSmarts("P([Cl,Br,I,F])([Cl,Br,I,
      F])[Cl,Br,I,F]"),
    "acyl_halide": Chem.MolFromSmarts("[CX3](=O)[Cl,Br,I,F]"),
    "imidoyl_halide": Chem.MolFromSmarts("[CX3](=[NX2])[Cl,Br,I,F]"),

    "thioacyl_halide": Chem.MolFromSmarts("[CX3](=S)[Cl,Br,I,F]"),
    "organolithium": Chem.MolFromSmarts("[#6][Li]"),
    "organomagnesium": Chem.MolFromSmarts("[#6][Mg]"),
    "organoaluminium": Chem.MolFromSmarts("[#6][Al]"),
    "organozinc": Chem.MolFromSmarts("[#6][Zn]"),
    "organomercury": Chem.MolFromSmarts("[#6][Hg]"),
    "pyrrolic_N": Chem.MolFromSmarts("[nH]"),
    "pyridinic_N": Chem.MolFromSmarts("n"),
    "aromatic_O": Chem.MolFromSmarts("o"),
    "aromatic_S": Chem.MolFromSmarts("s"),
    "primary_amine": Chem.MolFromSmarts("[NX3H2][#6]"),
    "secondary_amine": Chem.MolFromSmarts("[NX3H1]([#6])[#6]"),
    "tertiary_amine": Chem.MolFromSmarts("[NX3]([#6])([#6])[#6]"),
    "ammonium_cation": Chem.MolFromSmarts("[NX4+]"),
    "amine_oxide": Chem.MolFromSmarts("[NX3+][O-]"),
    "enamine": Chem.MolFromSmarts("[NX3][CX3]=[CX3]"),
    "hydroxylamine": Chem.MolFromSmarts("NO"),
    "hemiaminal": Chem.MolFromSmarts("[CX4]([OX2H])([NX3])[#6]"),
    "hemiaminal_ether": Chem.MolFromSmarts("[CX4]([OD2])([NX3])[#6]"
      ),
    "thioaminal": Chem.MolFromSmarts("[CX4]([SX2H])([NX3])[#6]"),
    "thioaminal_ether": Chem.MolFromSmarts("[CX4]([SX2])([NX3])[#6]"
      ),
    "aminal": Chem.MolFromSmarts("[CX4]([NX3])([NX3])[#6]"),
    "primary_ketimine": Chem.MolFromSmarts("[CX3](=[NX2H])[#6]"),
    "secondary_ketimine": Chem.MolFromSmarts("[CX3](=[NX2][#6])[#6]"
      ),
    "primary_aldimine": Chem.MolFromSmarts("[CX3H1](=[NX2H])[#6]"),
    "secondary_aldimine": Chem.MolFromSmarts("[CX3H1](=[NX2][#6])
      [#6]"),
    "amidine": Chem.MolFromSmarts("N=C(N)[#6]"),
    "guanidine": Chem.MolFromSmarts("N=C(N)N"),
    "ketoxime": Chem.MolFromSmarts("[CX3](=N[OX2H])[#6]"),
    "aldoxime": Chem.MolFromSmarts("[CX3H1](=N[OX2H])[#6]"),
    "hydrazone": Chem.MolFromSmarts("[NX3][NX2]=[CX3]"),
    "amide": Chem.MolFromSmarts("[NX3][CX3](=O)[#6]"),
    "amidate_anion": Chem.MolFromSmarts("[NX3][CX3](=O)[O-]"),
    "imide": Chem.MolFromSmarts("[CX3](=O)N[CX3](=O)"),
    "carbamic_acid": Chem.MolFromSmarts("NC(=O)O"),
    "carbamate_ester": Chem.MolFromSmarts("NC(=O)O[#6]"),
    "carbamate_anion": Chem.MolFromSmarts("[NX3][CX3](=O)[O-]"),
```

```
    "azide": Chem.MolFromSmarts("N=[N+]=[N-]"),
    "azo": Chem.MolFromSmarts("N=N"),
    "hydrazine": Chem.MolFromSmarts("NN"),
    "acylhydrazine": Chem.MolFromSmarts("[CX3](=O)NN"),
    "cyanate": Chem.MolFromSmarts("OC#N"),
    "isocyanate": Chem.MolFromSmarts("N=C=O"),
    "nitrile": Chem.MolFromSmarts("C#N"),
    "isonitrile": Chem.MolFromSmarts("N#C"),
    "cyanamide": Chem.MolFromSmarts("N=C=N"),
    "carbodiimide": Chem.MolFromSmarts("N=C=N"),
    "nitrate_ester": Chem.MolFromSmarts("O[N+](=O)[O-]"),
    "nitrite_ester": Chem.MolFromSmarts("ON=O"),
    "nitro": Chem.MolFromSmarts("[NX3](=O)=O"),
    "nitroso": Chem.MolFromSmarts("N=O"),
    "nitrosamine": Chem.MolFromSmarts("N(N)=O"),
    "iminium_cation": Chem.MolFromSmarts("[CX3]=[NX3+]"),
    "nitrone": Chem.MolFromSmarts("[NX3+](=O)[O-]"),
    "nitronic_acid": Chem.MolFromSmarts("[CX3](=N(=O)O)O"),
    "urea": Chem.MolFromSmarts("NC(=O)N"),
    "azoxy": Chem.MolFromSmarts("N(=O)N"),
    "N_oxoammonium": Chem.MolFromSmarts("[NX4+]=O"),
    "hydroxamic_acid": Chem.MolFromSmarts("NC(=O)NO"),
    "hydroxamate": Chem.MolFromSmarts("NC(=O)N[O-]"),
    "azanide": Chem.MolFromSmarts("[N-]"),
    "nitronate": Chem.MolFromSmarts("[O-][N+](=O)[#6]"),
    "mercaptan": Chem.MolFromSmarts("[SX2H]"),
    "thiolate_anion": Chem.MolFromSmarts("[S-][#6]"),
    "organic_sulfide": Chem.MolFromSmarts("[SX2]([#6])[#6]"),
    "thioenol": Chem.MolFromSmarts("C=C[SX2H]"),
    "enedithiol": Chem.MolFromSmarts("C=C([SX2H])[SX2H]"),
    "thioenolate": Chem.MolFromSmarts("C=C[S-]"),
    "thioenol_ether": Chem.MolFromSmarts("C=C[SX2][#6]"),
    "organic_disulfide": Chem.MolFromSmarts("[#16X2]-[#16X2]"),
    "sulfenic_acid": Chem.MolFromSmarts("[#16X2][OX2H]"),
    "sulfenic_ester": Chem.MolFromSmarts("[#16X2][OX2][#6]"),
    "sulfenamide": Chem.MolFromSmarts("[#16X2][NX3]"),
    "sulfoxide": Chem.MolFromSmarts("[#16X3](=O)"),
    "sulfone": Chem.MolFromSmarts("[#16X4](=O)(=O)"),
    "sulfinylamine": Chem.MolFromSmarts("S(=O)N"),
    "sulfinic_acid": Chem.MolFromSmarts("S(=O)[OX2H]"),
    "sulfonic_acid": Chem.MolFromSmarts("S(=O)(=O)[OX2H]"),
    "sulfinate_ester": Chem.MolFromSmarts("S(=O)(O[#6])[#6]"),
    "sulfonate_ester": Chem.MolFromSmarts("S(=O)(=O)O[#6]"),
    "sulfinate_anion": Chem.MolFromSmarts("S(=O)([O-])[#6]"),
    "sulfonate_anion": Chem.MolFromSmarts("S(=O)(=O)[O-]"),
    "thiosulfonate": Chem.MolFromSmarts("S(=O)(=O)SS"),
    "organosulfite": Chem.MolFromSmarts("S(=O)(O[#6])O[#6]"),
    "organosulfate": Chem.MolFromSmarts("S(=O)(=O)(O[#6])O[#6]"),
    "dialkylsulfates": Chem.MolFromSmarts("OS(=O)(=O)O[#6]"),
    "sulfinamide": Chem.MolFromSmarts("S(=O)N[#6]"),
    "sulfonamide": Chem.MolFromSmarts("S(=O)(=O)N"),
    "sulfamic_acid": Chem.MolFromSmarts("NS(=O)(=O)O"),
    "sulfamate": Chem.MolFromSmarts("NS(=O)(=O)O[#6]"),
    "sulfamide": Chem.MolFromSmarts("NS(=O)(=O)N"),
    "thiocyanate": Chem.MolFromSmarts("S-C#N"),
    "isothiocyanate": Chem.MolFromSmarts("N=C=S"),
    "thioketone": Chem.MolFromSmarts("C(=S)[#6]"),
    "thioketene": Chem.MolFromSmarts("C=C=S"),
    "thial": Chem.MolFromSmarts("[CX3H1](=S)[#6]"),
    "thioamide": Chem.MolFromSmarts("NC(=S)[#6]"),
    "thiourea": Chem.MolFromSmarts("NC(=S)N"),
    "hemithioketal": Chem.MolFromSmarts("[CX4]([SX2H])([SX2])[#6]"),
    "hemithioacetal": Chem.MolFromSmarts("[CX4H1]([SX2H])([SX2])[#6]
      "),
    "dithiohemiketal": Chem.MolFromSmarts("[CX4]([SX2H])([SX2H])[#6]
      "),
    "dithiohemiacetal": Chem.MolFromSmarts("[CX4H1]([SX2H])([SX2H])
      [#6]"),
    "monothioketal": Chem.MolFromSmarts("[CX4]([SX2])([OD2])[#6]"),
    "monothioacetal": Chem.MolFromSmarts("[CX4H1]([SX2])([OD2])[#6]"
      ),
    "dithioketal": Chem.MolFromSmarts("[CX4]([SX2])([SX2])[#6]"),
    "dithioacetal": Chem.MolFromSmarts("[CX4H1]([SX2])([SX2])[#6]"),
    "carbothioic_S_acid": Chem.MolFromSmarts("C(=S)S"),
    "carbothioic_O_acid": Chem.MolFromSmarts("C(=S)O"),
    "thiol_form_thiocarboxylate": Chem.MolFromSmarts("C(=O)[S-]"),
    "thione_form_thiocarboxylate": Chem.MolFromSmarts("C(=S)[O-]"),
    "thiolester": Chem.MolFromSmarts("C(=O)S[#6]"),
    "thionoester": Chem.MolFromSmarts("C(=S)O[#6]"),
    "carbodithioic_acid": Chem.MolFromSmarts("C(=S)SS"),
    "carbodithioic_anion": Chem.MolFromSmarts("C(=S)S[S-]"),
    "carbodithioic_ester": Chem.MolFromSmarts("C(=S)S[#6]"),
    "monothiocarbonate": Chem.MolFromSmarts("C(=O)(O)S"),
    "xanthic_acid": Chem.MolFromSmarts("C(=S)(O)S"),
    "xanthate": Chem.MolFromSmarts("C(=S)(O[#6])S[#6]"),
    "xanthate_anion": Chem.MolFromSmarts("C(=S)(O[#6])[S-]"),
    "dithiocarbonate": Chem.MolFromSmarts("C(=S)(S)S"),
    "O_thiocarbamic_acid": Chem.MolFromSmarts("NC(=S)O"),
    "S_thiocarbamic_acid": Chem.MolFromSmarts("NC(=O)S"),
    "O_thiocarbamate": Chem.MolFromSmarts("NC(=S)O[#6]"),
    "S_thiocarbamate": Chem.MolFromSmarts("NC(=O)S[#6]"),
    "O_thiocarbamate_anion": Chem.MolFromSmarts("NC(=S)[O-]"),
    "S_thiocarbamate_anion": Chem.MolFromSmarts("NC(=O)[S-]"),
    "thioimidic_acid": Chem.MolFromSmarts("C(=S)N"),
    "thioimidate_anion": Chem.MolFromSmarts("C(=S)[N-]"),
    "thioimidate": Chem.MolFromSmarts("C(=S)N[#6]"),
    "dithiocarbamic_acid": Chem.MolFromSmarts("NC(=S)S"),
    "dithiocarbamate": Chem.MolFromSmarts("NC(=S)S[#6]"),
    "dithiocarbamate_anion": Chem.MolFromSmarts("NC(=S)[S-]"),
    "sulfonium_ion": Chem.MolFromSmarts("[S+]"),
    "sulfonium_ylide": Chem.MolFromSmarts("[S+][C-]"),
    "sulfoxonium_ylide": Chem.MolFromSmarts("[S+](=O)[C-]"),
}
```

Listing 1: Full FG_SMARTS dictionary used for functional group matching.

A.2 Knowledge Graph Construction

The knowledge graph is implemented in Neo4j and organized into five node types connected by six edge types, forming the hierarchical structure illustrated in Figure 2. At the base layer, **Molecule** nodes store each compound's canonical SMILES and a hash-based identifier, while **Solvent** nodes (a separate label) represent molecules identified as non-reactive participants. Above this, **FunctionalGroup** nodes represent the 220 named substructure patterns, linked to molecules via HAS_FUNCTION_GROUP edges. At the reaction abstraction layer, **ReactionType** nodes encode ten coarse categories (e.g., Oxidation, Substitution), and **ReactionTypeId** nodes capture the 34 fine-grained transformation rules, connected by BELONGS_TO edges. Functional groups are linked to their associated reaction type IDs through PARTICIPATES_IN edges annotated with the specific functional group transformation pattern (e.g., "alcohol + carboxylic_acid $\rightarrow$ ester"), while solvents are linked via SOLVENT_IN edges. At the top layer, **Reaction** nodes represent individual reaction instances, each storing the atom-mapped reaction SMILES, reactant and product molecule identifiers, formed and consumed functional group lists, and a textual mechanistic explanation. Reaction nodes are connected to their corresponding ReactionType via CLASSIFIED_AS edges, and to participant Molecule and Solvent nodes via PARTICIPATES_IN and SOLVENT_IN edges respectively. All nodes and edges are created using idempotent `MERGE` statements, ensuring that the graph can be incrementally updated without duplication.

B Zero-shot Prompt in Evaluation

All the system prompt for model to do zero-shot reasoning is shown below:

VANILLA PROMPTS (without Knowledge Graph)

— Functional Group Identification —

[System Prompt] You are an expert organic chemist. Your task is to identify which functional group is contained in the given molecule (provided as SMILES notation).

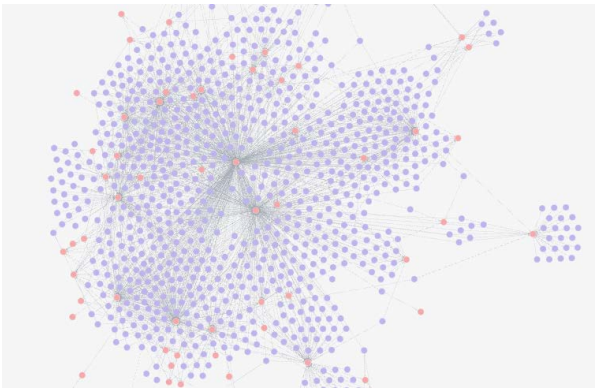


Figure 4: A sub-graph of 1000 relationships in the kG.

Think step by step: 1. First, analyze the SMILES string and identify the molecular structure 2. Identify the key functional groups present in the molecule 3. Compare with the given options 4. Select the correct answer

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

— Reaction Type Classification —

[System Prompt] You are an expert organic chemist. Your task is to identify the reaction type of a given chemical reaction (provided as SMILES notation).

Think step by step: 1. First, analyze the reactant and product SMILES strings and identify the molecular structures 2. Identify the key functional group transformations (which groups are consumed and which are formed) 3. Based on the transformations, determine the reaction type 4. Compare with the given options and select the correct answer

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

— Reaction Product Prediction —

[System Prompt] You are an expert organic chemist. Your task is to predict the product of a chemical reaction given the reactants and reaction type.

Think step by step: 1. First, analyze the reactant SMILES strings and identify the molecular structures 2. Consider the specified reaction type and what transformations it typically involves 3. Predict the product structure based on the reactants and reaction type 4. Compare with the given options and select the most likely product

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

— Reagent Prediction —

[System Prompt] You are an expert organic chemist. Your task is to predict the reagents (catalysts, solvents, or ancillary substances) needed for a given chemical reaction (provided as SMILES notation).

Think step by step: 1. First, analyze the reactant and product SMILES strings and identify the molecular structures 2. Identify the key functional group transformations (which groups are consumed and which are formed) 3. Based on the transformations, reason about what reagents, catalysts, or solvents would be needed 4. Compare with the given options and select the correct answer

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

KG-AUGMENTED PROMPTS (with Knowledge Graph)

— Functional Group Identification + KG —

[System Prompt] You are an expert organic chemist. Your task is to identify which functional group is contained in the given molecule.

You will be provided with: 1. The molecule in SMILES notation 2. Information retrieved from a Chemistry Knowledge Graph about the functional groups present in this molecule

Use the Knowledge Graph information to help you determine the correct answer. Think step by step: 1. Review the functional groups identified by the Knowledge Graph 2. Compare these with the given options 3. Select the correct answer based on the evidence

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

[Knowledge Graph Information] Molecule SMILES: smiles Source: source Functional groups found: comma-separated list

Please analyze the molecule and select the correct functional group from the options above.

— Reaction Type Classification + KG —

[System Prompt] You are an expert organic chemist. Your task is to identify the reaction type of a given chemical reaction (provided as SMILES notation).

You will be provided with: 1. The reaction in SMILES notation 2. Information retrieved from a Chemistry Knowledge Graph about functional group transformations or exact matches

Use the Knowledge Graph information to help you determine the correct reaction type. Think step by step: 1. Review the information from the Knowledge Graph (exact match or similar reactions) 2. Analyze the reactant and product structures and their functional group changes 3. Based on the transformations, determine the reaction type 4. Compare with the given options and select the correct answer

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

rag_context from get_reaction_type_rag()

Please analyze the reaction and select the correct reaction type from the options above.

— Reaction Product Prediction + KG —

[System Prompt] You are an expert organic chemist. Your task is to predict the product of a chemical reaction given the reactants and reaction type.

You will be provided with: 1. The reactants in SMILES notation and the reaction type 2. Information retrieved from a Chemistry Knowledge Graph about similar reactions or exact matches

Use the Knowledge Graph information to help you determine the correct product. Think step by step: 1. Review the information from the Knowledge Graph (exact match or similar reactions) 2. Analyze the reactant structures and reaction type 3. Consider the functional group transformations involved 4. Compare with the given options and select the most likely product

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

rag_context from get_reaction_product_rag()

Please analyze the reaction and select the correct product from the options above.

— Reagent Prediction + KG —

[System Prompt] You are an expert organic chemist. Your task is to predict the reagents (catalysts, solvents, or ancillary substances) needed for a given chemical reaction (provided as SMILES notation).

You will be provided with: 1. The reaction (reactants and product) in SMILES notation 2. Information retrieved from a Chemistry Knowledge Graph about exact matches or similar reactions with their reagents

Use the Knowledge Graph information to help you determine the correct reagent. Think step by step: 1. Review the information from the Knowledge Graph (exact match or similar reactions) 2. Analyze the reactant and product structures and the functional group changes 3. Based on the transformation, reason about what reagents would be needed 4. Compare with the given options and select the correct answer

You MUST end your response with your final answer in this exact format: **Answer: X** where X is A, B, C, or D.

[User Message] question

Options: options

rag_context from predict_reagent()

Please analyze the reaction and select the correct reagent from the options above.

MAX-EVAL-11: A Large-Scale Benchmark for Evaluating Large Language Models on Full-Spectrum ICD-11 Medical Coding

Ujjwal Singh
Max Healthcare
ujjwal.singh@maxhealthcare.com

Sarthak Deshwal
Plaid Inc.
sarthakdeshwal@duck.com

Nitish Dube
Max Healthcare
ntish.dube@maxhealthcare.com

Arjun Sharma
Max Healthcare
arjun.sharma@maxhealthcare.com

Abstract

The global transition to the ICD-11 taxonomy demands robust automated medical coding, yet comprehensive benchmarks to evaluate Large Language Models (LLMs) on this task remain absent. We introduce MAX-EVAL-11 [1] as a foundational resource, the first large-scale benchmark for full-spectrum ICD-11 medical coding. MAX-EVAL-11 comprises 10,000 MIMIC-III discharge summaries with mapped, expert-validated subset (12%) ICD-11 annotations spanning 99.87% of the diagnostic taxonomy. To better reflect clinical utility, we propose a novel hierarchical evaluation framework that assigns partial credit based on ICD-11's 5-level structure, addressing the brittleness of traditional exact-match metrics.

To illustrate benchmark utility, we evaluate LLMs on a stratified 500-patient sample, revealing significant performance gaps. The best-performing model (Claude 4 Sonnet) achieves a weighted score of 0.433, outperforming both general-purpose peers and specialized medical models (MedCoder). Crucially, all models exhibit near-zero exact match rates (0–4.8%) and rely primarily on hierarchical credit, underscoring the extreme difficulty of precise ICD-11 code generation. Furthermore, the superiority of general-purpose LLMs over legacy ICD-10 medical models (with ICD-11 codelist) suggests that broad reasoning capabilities currently outweigh domain-specific training for complex taxonomy scaling.

1 Introduction

Medical coding, transforming clinical narratives into standardized diagnostic codes is fundamental to modern healthcare for statistical analysis, reimbursement, and epidemiological research. The International Classification of Diseases (ICD), maintained by the WHO, provides the global standard for this task.

ICD-11, officially adopted in 2019 and effective since January 2022, represents a major evolution in medical classification (World Health Organization, 2019). As of 2024, 132 member states are implementing ICD-11, which introduces over 55,000 unique codes with post-coordination for granular clinical documentation (Harrison et al., 2021); (World Health Organization, 2024). Manual medical coding faces substantial challenges from growing documentation volumes, evolving regulations, and workforce shortages. The ICD-11 transition intensifies these pressures, requiring complex crosswalk mapping. Only 23.5% of ICD-9 codes achieve direct one-to-one ICD-11 mapping, complicating automated transition (Fung et al., 2021).

Recent AI advances, particularly large language models (LLMs), have sparked interest in automated medical coding. However, specialized models achieve F1 scores of only 0.539 to 0.719 on ICD-10 benchmarks (Baksi et al., 2025; Edin et al., 2023), while general-purpose LLMs face additional challenges (Hou et al., 2025). ICD-11's multidimensional taxonomy—capable of generating vast combinations via post-coordination—presents even greater complexity. Despite progress, existing benchmarks primarily focus on ICD-9 and ICD-10. MIMIC-III (Johnson et al., 2016) and MIMIC-IV-ICD (Nguyen et al., 2023) provide robust evaluation for legacy systems, but no comprehensive ICD-11 benchmark exists. Furthermore, traditional precision, recall, and F1 metrics inadequately capture the hierarchy of clinical relevance: accurate identification of a primary diagnosis is more critical than secondary conditions, yet existing metrics weight all predictions equally (Mullenbach et al., 2018).

We present **MAX-EVAL-11**, a comprehensive benchmark for evaluating LLM performance on ICD-11 medical coding. Our contributions are: **First**, a large-scale dataset with 10,000 real MIMIC-III discharge summaries with systematic

[1] Dataset: https://huggingface.co/datasets/mas-namtla/max-eval-11 (gated access).

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 282–291
July 3-4, 2026 ©2026 Association for Computational Linguistics

ICD-9 to ICD-11 conversion and complete taxonomy coverage. **Second**, a novel rank-weighted precision scoring methodology that assigns differential importance based on clinical relevance and diagnostic hierarchy, better reflecting real-world accuracy requirements. **Third**, baseline performance metrics using state-of-the-art LLMs (Claude, Gemini), demonstrating significant improvement opportunities in automated ICD-11 coding.

2 Related Work

2.1 Deep Learning Approaches for ICD Coding

Early neural approaches (RNNs, CNNs) established feasibility of automated coding but struggled with extreme multi-label classification (13,000+ codes) and long clinical documents (Choi et al., 2017; Baumel et al., 2018). Mullenbach et al. (2018) introduced CAML, using per-label attention to identify diagnosis-relevant text segments, establishing the label-wise attention paradigm for ICD coding. Subsequent refinements include LAAT (Vu et al., 2020), pseudo label-wise attention (Wu et al., 2021), and HiLAT (Liu et al., 2022), progressively improving efficiency and hierarchical coding accuracy. The integration of pretrained language models brought further improvements. Huang et al. (2022) introduced PLM-ICD, incorporating domain-specific pretraining, segment pooling, and specialized label attention, achieving 59.8% micro-F1 on MIMIC-III. Liu et al. (2023) extended these capabilities with extreme multi-label transformers adapted for long clinical narratives (ClinicalBIGBIRD), reaching 60.8% micro-F1. However, these advances remain constrained to ICD-9 and ICD-10 evaluation, with no comprehensive ICD-11 benchmarks available despite the classification's 2019 WHO adoption and ongoing global implementation across 132 member states. Wu et al. (2025) introduced MIMIC-SR-ICD11, a dataset derived from MIMIC-IV for narrative-based ICD-11 diagnosis, though limited to primary diagnosis only and covering a narrow subset of the full taxonomy without a hierarchical evaluation framework.

2.2 Evaluation Frameworks and Limitations

Kim et al. (2022) introduced AnEMIC, providing standardized preprocessing and evaluation protocols for MIMIC-III benchmarks—an important step toward reproducibility. However, three fundamental limitations persist across existing evaluation frameworks, directly motivating MAX-EVAL-11's development.

Metric mismatch with clinical hierarchy. Standard precision, recall, and F1 metrics treat all diagnostic codes uniformly despite differing clinical salience (Kim et al., 2022). This approach fails to capture medical classification systems' hierarchical nature. For instance, predicting "Infectious Disease" (ICD-11 chapter 1A00-1G9Z) when the ground truth is "Bacterial Pneumonia" (CA40.0) receives zero F1 credit, yet the model correctly identified the broad disease category (infection) and affected organ system (respiratory)—partial accuracy valuable in clinical practice where physicians can narrow diagnoses from correct disease families. Moreover, the distinction between primary and secondary diagnoses carries critical clinical importance that flat metrics ignore: primary diagnoses drive treatment decisions and resource allocation, making errors in primary identification far more consequential than missing secondary comorbidities (e.g., failing to identify acute myocardial infarction as primary could delay life-saving interventions). Existing metrics weight all predictions equally, obscuring these safety-critical distinctions.

The benchmark-practice performance gap. Recent evaluations reveal troubling disconnects between constrained benchmark performance and real-world capability. Hou et al. (2025) demonstrate that models achieving 97–99% exact match on simplified ICD-9 test sets collapse to 3.85–10.90% when evaluated on realistic clinical scenarios. This dramatic degradation stems from two factors: *extreme label spaces* where models must identify 10–15 relevant codes from 50,000+ candidates per clinical note (fundamentally different from constrained benchmarks limiting evaluation to hundreds of common codes), and *realistic multi-diagnosis complexity* in discharge summaries containing multiple diagnoses spanning organ systems, temporal progression (admission → complications → discharge), and complex interactions (e.g., diabetes exacerbating pneumonia recovery) that simplified benchmarks lack.

Absent ICD-11 coverage. Despite ICD-11's official WHO adoption in 2019 and implementation across 132 member states as of 2024 (World Health Organization, 2019), existing benchmarks remain focused on legacy ICD-9 (Johnson et al., 2016) and ICD-10 (Nguyen et al., 2023) classifications. The few ICD-11 pilot studies remain narrow in scope, typically evaluating models on small

taxonomy subsets (500–1,000 codes) or synthetic data lacking real clinical documentation complexity. No comprehensive benchmark spans ICD-11's full 55,000-code diagnostic taxonomy with authentic clinical text, creating critical evaluation gaps as healthcare systems worldwide transition to the new standard.

2.3 Positioning MAX-EVAL-11

MAX-EVAL-11 addresses these gaps through: (1) full-spectrum ICD-11 coverage across 10,000 MIMIC-III discharge summaries; (2) hierarchical evaluation metrics reflecting clinical diagnostic priority; and (3) comprehensive LLM baselines on realistic multi-diagnosis scenarios.

3 Methodology

3.1 ICD-9 to ICD-11 Mapping Framework

The transition from ICD-9 (≈13,000 codes) to ICD-11 (≈55,000 codes) represents a fundamental shift in medical classification granularity (World Health Organization, 2019). Given an ICD-9 code $c_9 \in \mathcal{C}_9$, we seek to identify the optimal ICD-11 code set $S_{11} \subseteq \mathcal{C}_{11}$ that preserves clinical semantics. Our hybrid approach combines semantic embedding-based candidate retrieval (efficient search space reduction from 55,000+ codes) with LLM-based contextual selection (clinical disambiguation of related codes).

3.2 Hybrid Semantic-LLM Mapping Pipeline

Problem Formulation. Let $\mathcal{D}_9 = \{(c_9^{(i)}, t_9^{(i)})\}_{i=1}^{N_9}$ and $\mathcal{D}_{11} = \{(c_{11}^{(j)}, t_{11}^{(j)})\}_{j=1}^{N_{11}}$ represent the ICD-9 and ICD-11 code sets, where c denotes the code identifier, t denotes the textual description, $N_9 \approx 13{,}000$, and $N_{11} \approx 55{,}000$. Our mapping function is:

$$f : \mathcal{C}_9 \to 2^{\mathcal{C}_{11}} \times [1, 10] \tag{1}$$

where $f(c_9) = (S_{11}, \sigma)$ returns ICD-11 codes S_{11} with LLM-assigned quality score $\sigma \in [1, 10]$.

Semantic Candidate Generation. While WHO provides partial ICD-9–ICD-11 crosswalk tables, these cover only 23.5% of ICD-9 codes with direct mappings (Fung et al., 2021); our hybrid pipeline addresses the remaining 76.5% for which no official mapping exists. We employ Bio_ClinicalBERT (Alsentzer et al., 2019), pretrained on 2 million MIMIC-III clinical notes, achieving F1 = 0.89 on i2b2 2010 medical entity recognition. Its 768-dimensional embeddings capture fine-grained semantic distinctions essential for differentiating clinically related diagnostic codes. We use the pretrained model without fine-tuning, as the task requires general medical semantic similarity rather than task-specific classification.

For each ICD-9 code, we construct composite inputs via string concatenation ($\oplus$):

$$x_9^{(i)} = c_9^{(i)} \oplus \text{“ - ”} \oplus t_9^{(i)} \tag{2}$$

and encode using Bio_ClinicalBERT's encoder ϕ : String $\to R^{768}$:

$$e_9^{(i)} = \phi(x_9^{(i)}) \tag{3}$$

Similarly, we encode all ICD-11 descriptions to obtain embeddings $e_{11}^{(j)}$. We identify top-$k = 10$ candidates via cosine similarity:

$$\text{sim}(e_9^{(i)}, e_{11}^{(j)}) = \frac{e_9^{(i)} \cdot e_{11}^{(j)}}{\|e_9^{(i)}\| \|e_{11}^{(j)}\|} \tag{4}$$

where $k = 10$ balances efficiency with coverage for one-to-many mappings common in ICD-9 to ICD-11 transitions.

LLM-Based Contextual Selection. Top-10 candidates often contain clinically related but distinct codes (e.g., bacterial vs. viral pneumonia) requiring medical reasoning to disambiguate. We employ Gemini 2.0 Flash, selected after comparing GPT-4, Claude 3.7, and Gemini 2.0 on 100 sample mappings: Gemini achieved 87% expert agreement versus 79% (GPT-4) and 82% (Claude 3.7). Its structured output capabilities and 1-million token context enable processing complete taxonomies. We formulate prompts:

$$\mathcal{P}(c_9, \mathcal{C}^{\text{cand}}) = [\text{Task}_{\text{instr}}, \text{Code}_{\text{ctx}}, \text{Output}_{\text{fmt}}] \tag{5}$$

where $\text{Task}_{\text{instr}}$ instructs the model to *rate the clinical relevance of each ICD-11 candidate for the given ICD-9 code on a scale of 1–10 (10 = exact semantic match; 1 = unrelated)*; Code_{ctx} includes ICD-9 code c_9 with the 10 ICD-11 candidates and their descriptions; and $\text{Output}_{\text{fmt}}$ specifies a JSON schema of `{code: string, score: int}` per candidate. The LLM returns $M_i = \{(c_j^{11}, \sigma_{ij})\}$ where $\sigma_{ij} \in [1, 10]$. We retain all candidates with $\sigma_{ij} \geq 7$ as the final mapping set S_{11}; if no candidate meets this threshold, the highest-scoring candidate is retained to ensure full coverage.

Methodological Independence. Label generation uses only ICD code taxonomies for code-to-code

translation (Gemini 2.0 Flash), while evaluation requires clinical narrative understanding from discharge summaries (Gemini 2.5 Flash), distinct models, tasks, and information sources (see Conclusion).

3.3 Quality Score-Based Stratification and Quality Assessment

We stratify by mapping quality score Σ_p, the average LLM-assigned score across patient p's mappings:

$$\Sigma_p = \frac{1}{|M_p|} \sum_{(c_9, c_{11}, \sigma) \in M_p} \sigma \tag{6}$$

where M_p represents all mappings for patient p. Binary stratification with $\tau = 7.0$ yields:

$$\text{Conf}(p) = \begin{cases} \text{High} & \text{if } \Sigma_p \geq 7.0 \\ \text{Low} & \text{if } \Sigma_p < 7.0 \end{cases} \tag{7}$$

This threshold was validated on 200 manual mappings: $\sigma \geq 7.0$ achieved 89% expert agreement versus 67% for $\sigma < 7.0$. The *High-Confidence Subset* ($\Sigma_p \geq 7.0$) optimizes for maximum reliability (billing, compliance), while *All-Matches Subset* enables robust evaluation across quality conditions.

We assess mapping quality via Coverage Rate $\rho = |\{p : |M_p| > 0\}|/N_p$ (proportion successfully mapped), Expansion Factor $\alpha = \sum_p |C_{11}^{(p)}| / \sum_p |C_9^{(p)}|$ (granularity increase, where $C_9^{(p)}$ and $C_{11}^{(p)}$ are code sets for patient p), and Semantic Preservation $\psi = \frac{1}{N_m} \sum_{i=1}^{N_m} \max_j \text{sim}(c_9^{(i)}, c_{11}^{(j)})$ (semantic alignment, where N_m is total mappings and $\text{sim}(\cdot, \cdot)$ is Equation 4). Values $\alpha > 1$ and $\psi > 0.75$ indicate successful ICD-11 expansion with semantic fidelity. The threshold $\psi > 0.75$ reflects the cosine similarity range at which BioClinicalBERT embeddings reliably preserve clinical semantics: values below 0.75 in our BioClinicalBERT embedding space correspond to pairs where the primary diagnostic category (e.g., infectious vs. cardiovascular) differs, as verified on the 200-mapping validation subset used for confidence threshold calibration (Section 3.3).

3.4 Hierarchical Evaluation Metrics

Traditional metrics assign binary credit (1 for exact match, 0 otherwise), ignoring ICD-11's hierarchical taxonomy. A model predicting "CA40.0Z Bacterial pneumonia, unspecified" for ground truth "CA40.0 Bacterial pneumonia" demonstrates clinical understanding despite technical mismatch. Similarly, primary diagnosis accuracy matters more than secondary coding, yet flat metrics weight equally. We propose multi-component scoring reflecting clinical priorities.

Hierarchical Credit Assignment. ICD-11's 5-level hierarchy (Chapter → Block → Category → Subcategory → Extension) enables partial credit $h(c_{\text{true}}, c_{\text{pred}})$ based on closest common ancestor:

$$h(c_{\text{true}}, c_{\text{pred}}) = \begin{cases} 1.0 & \text{exact} \\ 0.9 & \text{parent} \\ 0.8 & \text{grandparent} \\ 0.7 & \text{great-grandparent} \\ 0.6 & \text{chapter} \\ 0.0 & \text{unrelated} \end{cases} \tag{8}$$

Credits decrease uniformly by Δ=0.1 per taxonomy level from exact match (1.0) down to chapter-level match (0.6), reflecting ICD-11's 5-level hierarchy where each level represents a meaningful clinical distinction; codes sharing no chapter (e.g., an infectious disease predicted as a metabolic disorder) receive h=0.0 as they provide no clinically actionable signal regardless of proximity.

Metric Components. *Exact Match* (EM) measures precise coding:

$$\text{EM} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i \cap G_i|}{|P_i \cup G_i|} \tag{9}$$

where P_i and G_i are predicted and ground truth codes for patient i.

Clinical Precision (CP) weights by diagnostic priority:

$$\text{CP} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P_i|} \sum_{j=1}^{|P_i|} w_j \cdot 1[p_{ij} \in G_i] \tag{10}$$

where $1[\cdot]$ is the indicator function, $w_1 = 1.0$ (primary), $w_2 = 0.8$ (secondary), $w_{j \geq 3} = 0.6$ (comorbidities). E.g., for G={CA40.0, 5A10.0, 8B20.0}, P={CA40.0, 5A10.1, 1A00}: $CP = \frac{1.0+0+0}{3} = 0.33$.

Points Earned (PE) quantifies hierarchical partial credit:

$$\text{PE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in P_i} \max_{g \in G_i} h(p, g)}{|G_i|} \tag{11}$$

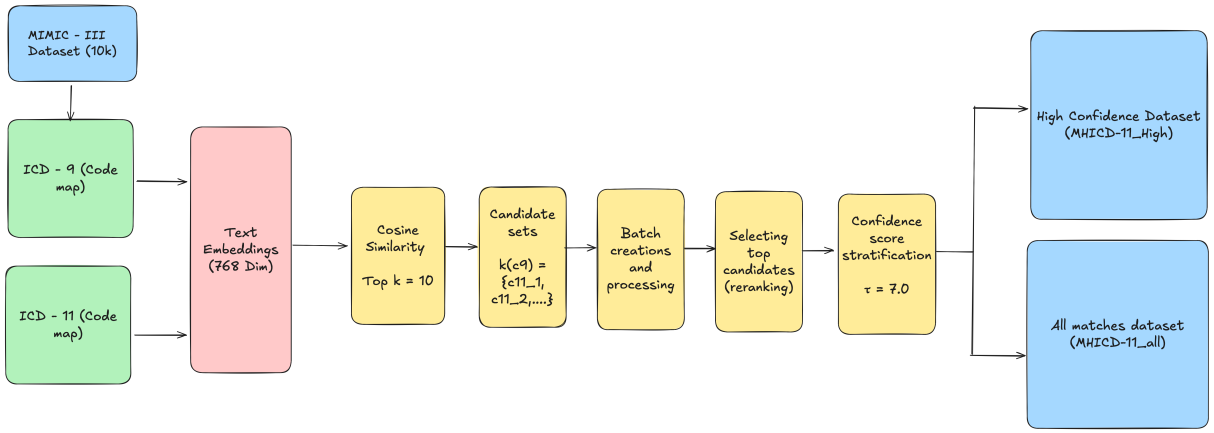


Figure 1: Overview of the hybrid semantic-LLM pipeline for ICD-9 to ICD-11 mapping. The process combines Bio_ClinicalBERT embeddings for candidate generation with Gemini 2.0 Flash for context-aware mapping selection, resulting in confidence-stratified benchmark datasets.

This rewards correct disease categories despite subtype errors (e.g., "CA40 Pneumonia, unspecified" for "CA40.0 Bacterial pneumonia" earns $h = 0.8$).

Hierarchical Match (HM) averages best alignments. For each prediction p, we find its best ground truth match: $\max_{g \in G_i} h(p, g)$. Averaging over predictions per patient, then across patients:

$$\text{HM} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P_i|} \sum_{p \in P_i} \max_{g \in G_i} h(p, g) \qquad (12)$$

Composite Weighted Score. We combine components with clinical weights:

$$\text{Score} = 0.5 \cdot \text{EM} + 0.3 \cdot \text{CP} + 0.15 \cdot \text{PE} + 0.05 \cdot \text{HM} \qquad (13)$$

Exact matching (0.5) prioritizes billing/regulatory accuracy; clinical precision (0.3) emphasizes primary diagnosis; points earned (0.15) rewards near-misses; hierarchical match (0.05) credits broad categorization. Unlike nDCG, our metric incorporates ICD-11-specific hierarchy and clinical priority weighting. Pilot validation (2 expert coders, 50 cases) showed Spearman ρ=0.79 vs. ρ=0.61 for F1.

4 Dataset Description

4.1 Source Data and Patient Selection

We derive our benchmark from MIMIC-III (Johnson et al., 2016), selecting 10,000 patients through stratified sampling to preserve clinical characteristic distributions while maintaining computational tractability. Each patient record contains discharge summaries with ICD-9 diagnosis codes. The dataset encompasses 6,984 unique ICD-9 codes across 118,106 total code instances (mean 11.81 codes per patient), representing comprehensive diagnostic coverage encountered in critical care settings. The ICD-9 distribution follows the characteristic heavy-tailed pattern of medical coding systems, with high variance indicating diverse diagnostic complexity.

4.2 ICD-9 to ICD-11 Mapping and Coverage

Our hybrid semantic-LLM pipeline (Section 3) transforms each patient's ICD-9 diagnosis set into corresponding ICD-11 codes. Patient-level coverage achieves 99.71% for high-confidence mappings ($\Sigma \geq 7.0$) and 99.87% overall, with code-level expansion factors of 1.099 (high-confidence) and 1.274 (all-matches). These metrics demonstrate successful expansion from ICD-9 to ICD-11's more granular classification, with each ICD-9 code mapping to an average of 1.1–1.3 ICD-11 codes, reflecting the enhanced diagnostic specificity of the newer taxonomy.

4.3 Benchmark Variants

We provide two benchmark datasets stratified by mapping confidence (Table 1). The high-confidence subset ($\Sigma \geq 7.0$) comprises 9,971 patients with 129,785 ICD-11 codes (mean 13.02 per patient), optimized for high-precision applications such as billing systems and regulatory compliance. The all-matches subset includes 9,987 patients with 150,502 ICD-11 codes (mean 15.07 per patient), providing comprehensive coverage for robust evaluation under diverse confidence conditions. Both variants include prescription data (drug names, dosages, timestamps) to support diagnosis-

treatment coherence analysis.

Metric	High-Conf.	All-Matches
Patients	9,971	9,987
ICD-11 codes	129,785	150,502
Mean codes/patient	13.02	15.07
Coverage	99.71%	99.87%
Expansion factor	1.099	1.274
Confidence threshold	≥ 7.0	All ≥ 1.0

Table 1: Benchmark dataset variant statistics.

4.4 Human Validation and Quality Control

To ensure benchmark reliability, we conducted manual validation with two clinical coders who independently reviewed 1,203 discharge summaries (12% of dataset) using stratified sampling across ICD-11 chapters. We also performed systematic quality analysis across all 10,000 records.

4.4.1 Confidence-Based Quality Tiers

Our validation reveals distinct quality tiers (Table 2). High-confidence mappings (85% of dataset) achieve expansion ratios of 0.8–1.5, indicating semantically consistent ICD-9→ICD-11 translation with complete coverage, suitable for production deployment. Low-confidence mappings (15%) exhibit lower expansion ratios and require manual review before clinical use.

Tier	%	Exp.	Utility
High ($\geq$7)	85	0.8–1.5	Production ready
Low (<7)	15	<0.8	Manual review

Table 2: Quality tiers by confidence threshold (σ). Expansion ratio = avg ICD-11 codes per ICD-9 code.

4.4.2 Domain-Specific Quality

Clinical domain analysis (Table 3) reveals variable mapping complexity. Domains with stable ICD-9→ICD-11 correspondences (infectious diseases, cardiovascular) achieve 89–92% success, while domains with structural taxonomy changes (external causes, procedures) show lower rates (38–45%), consistent with known ICD-11 architectural differences.

4.4.3 Error Analysis

Analysis of 273 disagreements (10.8% of validated codes) shows 70% stem from clinical ambiguity rather than pipeline failures: **(1)** Secondary diagnosis prioritization (42%): coding pre-existing conditions from history as active diagnoses—experts disagreed on 28% of these. **(2)** Granularity selection (28%): choosing specificity levels where both

Domain	Exp.	Succ.	Pattern
Infectious	1.2	92%	Excellent
Cardiovascular	1.1	89%	Consistent
Neoplasms	1.3	85%	Enhanced
Symptoms	0.7	62%	Variable
Ext. Causes	0.6	45%	Challenges
Procedures	0.4	38%	Manual

Table 3: Mapping quality by clinical domain (10K patients).

are valid. **(3)** Rare codes (18%): <10 instances in MIMIC-III, affecting only 2.3% of total data. **(4)** Cross-taxonomy inference (12%): extrapolating from ICD-9's coarser categories.

4.4.4 Benchmark Validity

Our validation supports three use cases: **(1)** High-precision evaluation using 85% subset (κ=0.67), comparable to ICD-10 inter-annotator agreement (Kim et al., 2022), for publication-quality comparison with near-expert labels. **(2)** Standard evaluation using full dataset ($\kappa = 0.52$) with quality comparable to existing benchmarks—MIMIC-III's ICD-9 labels contain documented errors (Kim et al., 2022) yet lack validation. **(3)** Confidence-aware training incorporating scores into loss functions. MAX-EVAL-11 provides transparent quality metrics exceeding existing ICD coding benchmarks.

4.5 Dataset Accessibility

Benchmark datasets are provided in CSV format with standardized schemas (SUBJECT_ID, HADM_ID, ICD9_CODES_LIST, ICD11_CODES_MAPPED, CONFIDENCE_SCORES, MEDICATIONS). All data adheres to MIMIC-III de-identification standards (Goldberger et al., 2000), and processing scripts are provided to ensure reproducibility.

5 Experimental Results

5.1 Baseline Model Performance

We evaluated three major categories of models on our ICD-11 prediction benchmark: Claude models (Anthropic), Gemini 2.5 Flash (Google), and MedCoder (specialized medical coding model – ICD-10) (Baksi et al., 2025). Table 4 presents the comprehensive performance evaluation across all baseline models. Claude 4 Sonnet achieves the highest overall performance with a final score of 0.433, demonstrating superior clinical precision (0.433) and coverage (0.353). Notably, Claude models consistently outperform both the specialized medical

model (MedCoder) and the multimodal Gemini model across all evaluation metrics. MedCoder (ICD-10 trained; Baksi et al., 2025) serves as a *zero-shot ICD-11 transfer baseline*, evaluated with the ICD-11 code list provided at inference without weight modification, representing the performance floor for domain-specific models without ICD-11 adaptation.

5.1.1 Evaluation Protocol

For each patient in the evaluation dataset, models receive:

- **Input**: Complete MIMIC-III discharge summary (chief complaint, history, medications, labs, hospital course, discharge diagnosis).
- **Task:** Generate ranked list of ICD-11 diagnosis codes based solely on clinical narrative
- **Constraints:** Separate pipeline, no access to original ICD-9 codes or mapping or ICD code taxonomies

The evaluation simulates real-world clinical coding scenarios where human coders or AI systems must assign ICD-11 codes directly from clinical documentation without reference to legacy coding systems. Models are prompted with standardized instructions requesting comprehensive ICD-11 code prediction with confidence scores and clinical justification. All models receive identical inputs to ensure fair comparison across architectures and capabilities. The prompt is: *Given the discharge summary, identify all diagnoses and provide their ICD-11 codes as a ranked, comma-separated list ordered by clinical priority.* Evaluation was conducted on a stratified random sample of n=500 patients, selected to reflect the full distribution of ICD-11 chapters and diagnosis complexity; Full-benchmark evaluation (n=10,000) is left for future work contingent on compute resources; we report stratified sampling to ensure chapter-level representativeness of the n=500 subset.

5.2 Statistical Analysis

Table 5 presents the statistical significance analysis using Wilcoxon signed-rank tests, revealing substantial performance gaps between model categories.

5.3 Key Findings

MedCoder serves as a zero-shot ICD-10→ICD-11 transfer baseline and is not a peer competitor. It establishes the performance floor for domain-specific models without ICD-11 adaptation. Claude models achieve superior performance (35.8% mean final score) compared to the domain-specific MedCoder (24.5% final score), suggesting that large-scale pretraining and reasoning capabilities may be more valuable than specialized medical training for ICD-11 prediction tasks.

Exact matching remains challenging: All models show relatively low exact match rates (0-4.8%), indicating the inherent difficulty of precise ICD-11 code generation. Claude models demonstrate the strongest exact matching capabilities among evaluated systems.

Hierarchical understanding varies significantly: Claude models excel in hierarchical matching (19.4-40.2%), leveraging the taxonomic structure of ICD-11, while MedCoder shows limited hierarchical understanding (18.4%).

RAG integration challenges: Retrieval-augmented variants underperform their baselines (Claude RAG: 0.253 vs. 0.425; RAG+Reranker: 0.282 vs. 0.425), suggesting that naive RAG introduces noise rather than beneficial context. We implement two variants over a Qdrant vector index of all 55,000 ICD-11 code descriptions: **Claude RAG** appends retrieved codes to the prompt before generation; **Claude 4 RAG+Reranker** additionally reranks by semantic similarity before injection; both operate agentically, with the LLM issuing retrieval queries during inference. The performance degradation likely stems from semantic proximity confounding clinical specificity: when a patient presents with *left-sided heart failure*, retrieval surfaces semantically adjacent codes (e.g., right-sided heart failure, cardiomyopathy) that share embedding space but are clinically distinct, anchoring the model toward incorrect codes.

6 Conclusion

This work presents the first comprehensive benchmark for ICD-11 prediction using a large-scale clinical dataset, contributing three key advances to medical informatics research.

Methodological Contribution: We develop a novel hybrid semantic-LLM pipeline for automated ICD-9 to ICD-11 mapping, combining BioClinicalBERT embeddings with Gemini 2.0 Flash for context-aware medical code translation. Our approach achieves 99.87% mapping coverage across 10,000 patients, establishing a robust foundation

Table 4: Comprehensive Performance Evaluation of Baseline Models on ICD-11 Prediction Task

Model	Final Score	Clinical Precision	Exact Match	Hierarchical Match	Coverage Score	Samples (n)
Claude 4 Sonnet	**0.433**	**0.433**	**0.047**	0.375	**0.353**	500
Claude 3.5 Haiku (Baseline)	0.425	0.426	0.034	**0.402**	0.315	500
Claude 3.7 Sonnet	0.396	0.372	0.048	0.325	0.332	500
Gemini 2.5 Flash	0.341	0.315	0.016	0.286	0.330	500
Claude 4 RAG+Reranker	0.282	0.232	0.010	0.202	0.175	500
Claude 4 RAG	0.253	0.215	0.007	0.194	0.160	500
MedCoder	0.245	0.191	0.000	0.184	0.092	500

Table 5: The larger performance gap for Gemini vs. MedCoder (0.096) yields a higher p-value than Claude vs. Gemini (0.092) because Wilcoxon signed-rank tests reflect score consistency across patients, not only mean differences; MedCoder exhibits higher per-patient variance (near-zero scores on complex multi-diagnosis cases), reducing the rank-based significance despite the larger mean gap.

Comparison	Performance Gap	Relative Improvement	p-value
Claude-4-Sonnet vs Gemini-2.5-Flash	0.092	27.0%	$< 0.001^{***}$
Gemini-2.5-Flash vs MedCoder	0.096	39.2%	$< 0.01^{**}$
Claude-4-Sonnet vs MedCoder	0.188	76.7%	$< 0.001^{***}$
Claude Category vs Gemini Category	0.061	18.8%	$< 0.01^{**}$
Claude Category vs MedCoder Category	0.150	61.2%	$< 0.001^{***}$

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

for ICD-11 adoption.

Evaluation Framework: Our hierarchical metric assigns partial credit across ICD-11's 5-level taxonomy, achieving Spearman ρ=0.79 alignment with expert clinical judgment versus ρ=0.61 for standard F1.

Empirical Insights: General-purpose LLMs (Claude 4 Sonnet: 43.3% final score) meaningfully outperform specialized medical models (MedCoder: 24.5% final score) in zero-shot ICD-11 transfer, suggesting broad reasoning capabilities currently outweigh domain-specific ICD-10 training for taxonomy scaling, a finding reinforced by MedCoder's 0% exact match rate despite being purpose-built for medical coding. The benchmark datasets and evaluation framework are made publicly available to support continued research; Future work should focus on improving exact match performance, hierarchically-aware RAG integration, validating across diverse care settings, and exploiting the prescription data included in the benchmark but unused in this evaluation.

Methodological Independence. A potential concern is whether the use of Gemini 2.0 Flash in label generation creates bias toward Gemini models during evaluation. However, three factors demonstrate this is not the case:

(1) Task Difference: Label generation uses structured code taxonomies (code-to-code translation, while evaluation requires clinical text understanding (narrative-to-code generation)—fundamentally different capabilities requiring distinct model strengths.

(2) Information Sources: Mapping uses only ICD code descriptions/titles of ICD-9 and ICD-11 code tables; evaluation uses complete discharge summaries with clinical context, medications, procedures, temporal progression, and physician reasoning—information never seen during label creation.

(3) Empirical Evidence: Gemini 2.5 Flash achieves the lowest performance among major models (0.341 final score vs. Claude 4 Sonnet's 0.433, a 27% gap, Table 4), demonstrating no preferential advantage from its predecessor's involvement in label creation. If systematic bias existed, Gemini models would outperform rather than underperform. This performance gap actually validates that clinical coding ability (extracting diagnoses from narratives) is independent from taxonomy knowledge (code mapping). The substantial performance difference between Claude and Gemini models (statistically significant at $p < 0.001$, Table 5) reflects genuine differences in clinical

reasoning and medical text understanding capabilities, not artifacts of the label generation process; nonetheless, scores reflect benchmark-relative alignment rather than clinical deployment readiness.

7 Ethical Considerations

7.1 Data Privacy and Security

All patient data used in this study is sourced from the publicly available MIMIC-III database, which has undergone rigorous de-identification procedures following HIPAA Safe Harbor provisions (Goldberger et al., 2000). No additional patient identifiers were created during our mapping process, and all generated benchmarks maintain the original de-identification standards.

7.2 Bias and Fairness

The MIMIC-III dataset reflects the patient population of a single academic medical center, potentially introducing demographic and socioeconomic biases into our benchmark. Furthermore, MIMIC-III captures exclusively ICU admissions, where patients exhibit atypically high diagnostic complexity (mean 11.8 codes/patient); benchmark performance may not generalize to community hospital or outpatient coding scenarios with simpler, fewer diagnoses. The ICD-9 codes in the original dataset may also exhibit coding practices specific to the source institution, which could propagate through our ICD-11 mappings. Future work should validate across diverse care settings and patient populations.

7.3 Clinical Safety Implications

Automated ICD coding systems carry significant clinical and financial implications. Coding errors can affect patient care continuity, insurance reimbursements, and population health analytics. Our evaluation reveals that current models require human oversight, with exact match rates below 5% across all systems. Healthcare organizations should implement appropriate safeguards and validation procedures before deploying automated coding systems.

7.4 Algorithmic Accountability

The hybrid semantic-LLM approach introduces opacity in the mapping process, particularly in the LLM decision-making component. While we provide confidence scores and mapping justifications, the full reasoning process remains partially opaque. This limitation necessitates careful validation and monitoring in clinical deployment scenarios.

7.5 LLM Service Compliance

In accordance with PhysioNet's responsible use policy for credentialed datasets (PhysioNet, 2025), all cloud-based LLM processing of MIMIC-III discharge summaries was conducted exclusively through enterprise-tier services: Claude models via **Amazon Bedrock** (AWS enterprise account) and Gemini models via **Google Vertex AI** (GCP enterprise account), both explicitly listed as compliant services by PhysioNet (PhysioNet, 2025). Both platforms guarantee zero data retention, no use of submitted data for model training, and no human review of submitted content, satisfying all PhysioNet DUA requirements for processing credentialed health data with external LLM services.

8 Bibliographical References

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Krishanu Das Baksi, Elijah Soba, John J. Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack I. Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, and Sanmitra Bhattacharya. 2025. MedCodER: A generative AI assistant for medical coding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 449–459, Albuquerque, New Mexico. Association for Computational Linguistics.

Tal Baumel, Julien Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: Case study on ICD code assignment. In *AAAI Workshop on Health Intelligence*, New Orleans, LA. AAAI Press.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. pages 785–794, San Francisco, CA. ACM.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, pages 2572–2582. ACM.

Kin Wah Fung, Julia Xu, Shannon McConnell-Lamptey, Donna Pickett, and Olivier Bodenreider. 2021. Feasibility of replacing the ICD-10-CM with the ICD-11 for morbidity coding: A content analysis. *Journal of the American Medical Informatics Association*, 28(11):2404–2411.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Janine E. Harrison, Saskia Weber, Robert Jakob, and Christopher G. Chute. 2021. ICD-11: An international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(Suppl 6):206.

Z. Hou, M. Zhang, C. Shi, H. Shen, Y. Wang, T. Zhuang, and B. Liu. 2025. Enhancing medical coding efficiency through domain adaptation of language models. *NPJ Digital Medicine*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Juyong Kim, Abheesht Sharma, Suhas Shanbhogue, Jeremy C. Weiss, and Pradeep Ravikumar. 2022. AnEMIC: A framework for benchmarking ICD coding models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, UAE. Association for Computational Linguistics.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics*, 133:104161.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated ICD coding using extreme multi-label long text transformer-based models. *Artificial Intelligence in Medicine*, 144:102662.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh Tung Nguyen, Viktor Schlegel, Maurice Mulvenna, Raymond Bond, Lixin Su, Damien Coyle, and William O'Brien. 2023. MIMIC-IV-ICD: A new benchmark for extreme multilabel classification of ICD codes. *arXiv preprint arXiv:2304.13998*.

PhysioNet. 2025. Use of MIMIC data with large language models and online services.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 3335–3341. IJCAI.

World Health Organization. 2019. ICD-11: International classification of diseases 11th revision. `https://icd.who.int/en`.

World Health Organization. 2024. ICD-11 implementation progress report. Technical report, World Health Organization.

Y. Wu, Y. Yu, M. Zeng, and M. Li. 2021. A pseudo label-wise attention network for automatic ICD coding. *arXiv preprint arXiv:2106.06822*.

Yuexin Wu, Shiqi Wang, and Vasile Rus. 2025. Mimic-sr-icd11: A dataset for narrative-based diagnosis. In *Findings of Machine Learning for Health (ML4H)*.

Trustworthy NLP for Low-Resource Languages: Agent-Based Uncertainty Modeling for Hebrew Radiology Report Structuring

Hadas Ben-Atya[1], Naama Gavrielov[1], Zvi Badash[1], Gili Focht[2], Ruth Cytter-Kuint[2], Talar Hagopian[2], Dan Turner[2], Moti Freiman[1]
[1]Faculty of Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, Israel
[2]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, Jerusalem, Israel
Correspondence: hds@campus.technion.ac.il

Abstract

Reliable extraction of structured information from radiology reports using Large Language Models (LLMs) remains a significant challenge, particularly for complex, non-English texts such as Hebrew. This study proposes an agent-based, uncertainty-aware framework to enhance the reliability and interpretability of LLM predictions in clinical contexts. A total of 9,683 Hebrew radiology reports from Crohn's disease patients (2010–2023) across three medical centers were analyzed. Of these, 512 reports were manually annotated for six gastrointestinal organs and 15 pathological findings, while the remainder were automatically labeled using HSMP-BERT. Structured data extraction was performed with Llama 3.1 (Llama 3-8b-instruct) employing Bayesian Prompt Ensembles (BayesPE), which utilized six semantically equivalent prompts to quantify uncertainty. An Agent-Based Decision Model aggregated prompt outputs into five calibrated confidence levels and was benchmarked against three entropy-based approaches. Model performance was assessed using accuracy, F1 score, precision, recall, and Cohen's Kappa before and after filtering high-uncertainty cases. The agent-based model outperformed all baselines, achieving an F1 score of 0.3967, recall of 0.6437, and Kappa of 0.3006; after excluding cases with uncertainty $\geq$0.5, the F1 score increased to 0.4787 and Kappa to 0.4258. The proposed framework improves uncertainty calibration and predictive reliability, advancing the safe deployment of LLMs in medical data extraction.

1 Introduction

Radiology reports contain critical clinical information, essential for healthcare decision-making, retrospective studies, and radiologic image annotation. Despite ongoing efforts to implement structured reporting (Nobel et al., 2022), unstructured free-text reports remain dominant, posing challenges for large-scale data analysis due to their variability and lack of standardization. Automating the extraction of structured data addresses these issues by streamlining workflows, reducing manual effort, and improving data consistency and accessibility. This approach not only enhances patient care but also enables large-scale research through comprehensive meta-analyses, accelerated scientific discoveries, and stronger evidence-based practice (Hassanpour and Langlotz, 2016; Haendel et al., 2018; Langlotz et al., 2019).

Large language models (LLMs) can automate structured data extraction from radiology reports by processing unstructured text to efficiently yield clinically relevant insights (Busch et al., 2024; Doshi et al., 2024; Reichenpfader et al., 2024). However, privacy concerns around sensitive medical data have prompted interest in open-source LLMs, which offer greater transparency, adaptability, and control. Models like Llama 3.1 can extract meaningful insights from medical texts while mitigating risks associated with proprietary systems (Nowak et al., 2025).

Despite their potential, applying LLMs to radiology remains challenging, particularly for decision-critical tasks. A key limitation is their tendency to produce overconfident predictions, even when uncertain or incorrect, undermining trust in high-stakes medical settings (Shmidman et al., 2024; Tessler et al., 2024; Chen et al., 2024). Although uncertainty quantification methods have improved reliability, they exhibit unique structural limitations. Confidence elimination (CE) relies on explicit self-evaluation prompts but often suffers from sycophancy and poor calibration (Xiong et al., 2024). Token-level probabilities (TLP) utilize raw token distributions but are heavily restricted by proprietary API boundaries (Huang et al., 2025). Sample consistency (SC) measures generation diversity via temperature sampling but treats all outputs uniformly without weighting (Kompa et al., 2021).

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 292–311
July 3-4, 2026 ©2026 Association for Computational Linguistics

They often treat all responses equally, potentially ignoring variations in prompt quality or contextual relevance (Savage et al., 2025). These limitations are further amplified in underrepresented languages like Hebrew, where limited annotated data and linguistic differences hinder LLM adaptability. Addressing both overconfidence and linguistic gaps is essential for robust and equitable deployment across diverse clinical settings.

Bayesian Prompt Ensembles (BayesPE) (Tonolini et al., 2024) improve uncertainty estimation in LLM by combining semantically equivalent prompts and optimizing their weights by variational inference, applying these weights directly to the prompt-specific outputs to improve prediction reliability and calibration. However, reliance on a parametric model for prompt weights may limit flexibility in complex or dynamic settings.

Agent-based approaches (Xi et al., 2023) offer adaptive decision making for LLM outputs, potentially addressing BayesPE's limitations by incorporating more sophisticated uncertainty estimation processes.

1.1 Related Work

The application of AI in clinical text processing has advanced rapidly, particularly in structuring clinical narratives (Hassanpour and Langlotz, 2016). While large language models demonstrate high proficiency in English medical domains (Langlotz et al., 2019), their performance declines in low-resource and morphologically rich languages like Hebrew (Shmidman et al., 2024). To bridge this gap, recent paradigms explore agentic AI frameworks where multi-agent systems coordinate to solve complex clinical reasoning tasks (Zeng et al., 2024). However, utilizing these adaptive frameworks specifically for calibrating predictive uncertainty in low-resource medical processing remains significantly underexplored.

This study introduces a novel framework to address these limitations. This study aims, therefore, to introduce and evaluate an agent-based method for the extraction of structured data. Our approach utilizes uncertainty-aware LLM-based data from real-world Hebrew radiology reports from Crohn's disease patients. Finally, we compare its performance and robustness with BayesPE's probabilistic aggregation methods.

Contributions

- We introduce and evaluate a generalizable agent-based approach for quantifying uncertainty in Large Language Models (LLMs) used for structured data extraction in high-stakes clinical settings, specifically utilizing low-resource Hebrew radiology reports.
- We design an Agent Decision Model (using Llama 3 - 70B) that synthesizes outputs and explanations from a Bayesian Prompt Ensemble (BayesPE) and categorizes prediction confidence into five distinct levels (e.g., Definitely Yes/No, Likely Yes/No, Uncertain).
- We demonstrate that the agent-based approach provides the best balance of performance metrics (F1=0.3967, Cohen's Kappa=0.3006) compared to single-prompt baselines and entropy-based aggregation models.
- We validate the superior calibration of the Agent model, showing clearer separation between correct predictions (median uncertainty 0.0) and incorrect predictions, and confirm that filtering high-uncertainty cases significantly improves reliability.

2 Materials and methods

2.1 Data Collection

The retrospective multicenter study was approved by the institutional review board, with a waiver of informed consent due to its retrospective nature and the de-identification of data. Free-text Hebrew radiology reports of Crohn's disease patients were obtained from the *epi-IIRN* national inflammatory bowel disease (IBD) study cohort (Friedman et al., 2018), a validated database encompassing all four Israeli health maintenance organizations (HMOs), which collectively cover over 98% of the population. The dataset comprises **9,683 radiology reports**, each corresponding to an individual patient visit, derived from **8,093 unique patients** across three medical institutions. Each report documents a *magnetic resonance enterography* (MRE) or *computed tomography enterography* (CTE) examination performed between **2010 and 2023**.

2.1.1 Manual Annotation of Organ–Finding Pairs

A uniformly distributed random subset of **512 reports** was selected for manual annotation by an expert radiologist (T.H.). While relying on a single expert annotator presents a structural limitation regarding inter-annotator variability and potential subjective bias, it reflects the high cost and scarcity of specialized medical expertise. Each report was labeled for **six gastrointestinal organs** (Jejunum, Ileum, Cecum, Colon, Sigmoid, and Rectum) and

up to **15 pathological findings** per organ (e.g., inflammation, wall thickening, ulceration) (Bruining et al., 2018), yielding a total of **90 possible organ–finding combinations**.

Each combination was assigned one of four categorical labels: 1 = finding present (positive), 0 = finding absent (negative), 2 = organ resected (post-surgical), or 9 = organ not visible. For the main analysis, labels 2 and 9 were treated as negative, simplifying the task to a **binary classification** (finding present vs. absent). To ensure sufficient representation of positive cases, we excluded organ–finding pairs with $\leq$15 positive examples, resulting in **23 final classification targets**, which represent specific valid clinical pairings of a target gastrointestinal organ and an associated pathological condition (see Figure 4 in Appendix A).

2.1.2 NLP Based Annotations

We used HSMP-BERT, an internal BERT-based NLP model for structured data extraction from Hebrew radiology reports from Crohn's disease patients (Hazan et al., 2024a,b), to annotate the entire data set of 9,683 reports. The HSMP-BERT architecture was initialized from a pre-trained multilingual BERT checkpoint and fine-tuned using a masked language modeling objective followed by a multi-label classification head, optimized via AdamW with a learning rate of 2×10^{-5} and a batch size of 16 on the manually annotated training partition. The model was trained in the manually annotated subset, focusing on 23 common combinations of organ-finding. Performance metrics for these labels appear in Table 5 in the supplementary materials. To ensure high-quality data for BayesPE-based methods, we retained only labels with a Cohen's Kappa score greater than 0.7, resulting in eight final labels: Sign of ileum comb, inflammation of the ileum, pre-stenotic dilation of the ileum, stenosis of the ileum, enhancement of the ileum wall, thickness of the ileum wall, thickness of the rectum wall and sign of the sigmoid comb.

2.2 Large Language Model (LLM) Utilization

We used the Llama 3.1 model (Llama 3-8b-instruct) (AI@Meta, 2024), part of Meta's multilingual LLM collection, to extract structured data from Hebrew radiology reports. Although it supports several languages, Hebrew is not among them. The selection of Llama 3.1, despite its lack of official Hebrew optimization compared to models like Phi-4, Aya, or local Hebrew variants, was driven by its superior architectural scaling, high cross-lingual transferability, and widespread open-source deployment capability in secure clinical environments. Llama 3.1 utilizes an auto-regressive transformer architecture (Vaswani et al., 2017), further refined via supervised fine-tuning (SFT) (Sun, 2024) and reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2022; Ouyang et al., 2022) to better align with human preferences. Given its limited Hebrew support, we carefully adapted the model for this context, applying distinct prompt structures to evaluate cross-prompt variability. We prompted the model with various organ–finding combinations to extract structured data from Hebrew radiology reports, asking whether a specific finding was indicated. We also included a self-explanation step for each answer, aiding model validation and improving accuracy (see Appendix C for the full prompt schema).

2.3 Uncertainty Estimation with Bayesian Prompt Ensembles

We employed the Bayesian Prompt Ensembles (BayesPE) method (Tonolini et al., 2024) to estimate uncertainty in LLM-generated predictions by leveraging an ensemble of semantically equivalent prompts, without modifying the LLM's architecture or requiring retraining. In alignment with findings from (Tonolini et al., 2024), we selected six prompts to balance the benefits of prompt diversity with computational efficiency, as adding more prompts provided only marginal gains. We used ChatGPT (OpenAI, 2024) to design these six prompts, each with identical semantic intent to determine the presence of specific findings in radiology reports.

We developed an agent-based method and three entropy-based approaches to quantify model uncertainty using outputs from multiple prompts. Fig. 1 illustrates the different approaches we utilized to assess uncertainty.

2.4 Agent Decision Model

We propose an Agent Decision Model, functioning as a second-stage LLM component, to consolidate the outputs of multiple prompts, derive a final decision, and quantify its uncertainty. This model synthesizes responses and explanations from an ensemble of prompts, producing a unified decision with a rationale for its conclusion. The agent evaluates response consistency, assesses the clarity and coherence of explanations, and identifies indi-

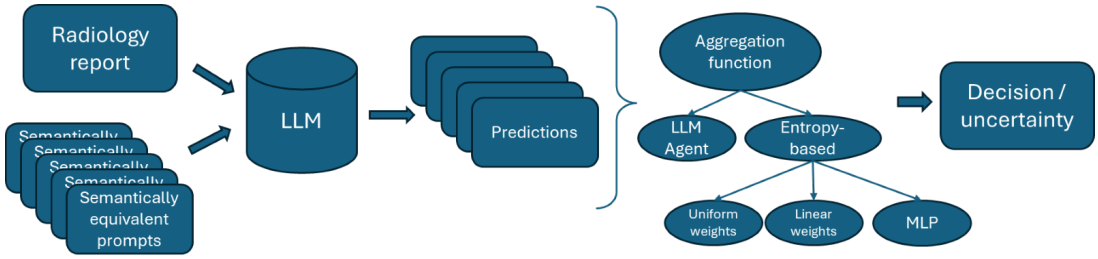


Figure 1: Illustration of the proposed Bayesian Prompt Ensemble pipeline for uncertainty-aware predictions. A radiology report and multiple semantically equivalent prompts are fed into an LLM, generating one prediction per prompt. These predictions are then aggregated to yield a final decision and uncertainty estimation. Note that the listed aggregation methods represent standalone, alternative pathways evaluated in isolation rather than interacting concurrent components.

cations of uncertainty. Based on these factors, it categorizes the final decision into one of five distinct confidence levels. This 5-level discretization was chosen to mimic clinical reporting conventions (e.g., highly likely, indeterminate) and to separate explicit certainty from qualified or ambiguous findings:

- Definitely Yes
- Likely Yes
- Uncertain
- Likely No
- Definitely No

The agent receives as input a set of answers ("Yes" or "No") and their corresponding explanations, generated by prompting a smaller, efficient LLM using multiple semantically equivalent questions. This design choice was motivated by two key factors: (1) asking the same question multiple times is significantly more cost- and time-efficient with smaller models; and (2) the inputs of this stage do not contain personally identifiable medical information, allowing the aggregation step to safely utilize more powerful models—either offline or online—without posing privacy risks. The agent evaluates these inputs based on the consistency of the answers, the clarity of their explanations, and the degree of ambiguity to produce its final decision.

To ensure interpretability and structure, the agent outputs its decision and reasoning in a predefined JSON format. In our implementation, we utilized the Llama 3 - 70B model as the agent.

The agent's uncertainty level is derived based on the confidence category of its decision:

- Definitely Yes/No: Uncertainty = 0
- Likely Yes/No: Uncertainty = 0.5
- Uncertain: Uncertainty = 1

This semi-quantitative approach enables the seamless integration of uncertainty measures into downstream analyses, enhancing the interpretability of the decision-making process (see Appendix C for detailed inputs/outputs). For the "Uncertain" response, a final "Yes" or "No" decision is determined by aggregating the probabilities for each option from the previous step and selecting the option with the highest overall score.

2.5 Entropy-based Decision Models

Alongside our agent-based decision model, we implemented three entropy-based models for comparison. These models quantify uncertainty using entropy, a measure of the disorder or unpredictability within the ensemble of prompt outputs. Higher entropy indicates greater disagreement among prompts, suggesting higher uncertainty in the final prediction. The final prediction and its entropy are determined by applying different weighting methods to the prompt outputs, as follows:

2.5.1 Uniform Weights

In this approach, all prompts are assigned equal weights: $w_i = 1/N$, where N is the number of prompts. This method operates under the assumption of no prior knowledge regarding the efficacy of individual prompts, thus distributing trust uniformly.

2.5.2 Linearly Optimized Weights

In this approach, we optimized the weights assigned to different prompts using a small validation set. We determine the weights by minimizing the following objective function:

$$\mathcal{L}_{\text{opt}} = \sum_{i=1}^{N} w_i \log P(y^*|a_i, x) - \lambda \sum_{i=1}^{N} w_i \log(w_i) \quad (1)$$

where y^* denotes the correct answer, a_i represents the i^{th} prompt structure, and x is the radiology report. The weights w_i are computed by applying a softmax transformation to the raw weights $w_{\text{raw},i}$. The first term represents the log-likelihood of the validation data, while the second term acts as an entropy-based regularizer, discouraging overconfidence in a single prompt. We performed linear optimization of prompt weights for each organ-finding label using a fixed subset of 50 samples from the manually annotated dataset.

2.5.3 Learnable Weights Using MLP

The uniform and linearly optimized weighting methods assume fixed or simple relationships between prompts, limiting their ability to capture complex interactions. To address this, we introduce a Multi-Layer Perceptron (MLP) for adaptive prompt weight optimization. The MLP dynamically adjusts weights based on prompt behavior across the dataset, enabling more accurate weighting in scenarios with intricate dependencies between prompts. The forward pass for the model is defined as follows:

$$\mathbf{w} = \text{softmax}(\text{MLP}(\mathbf{h})) \tag{2}$$

where $\mathbf{h}$ represents the fixed prompt embeddings, and $\mathbf{w}$ are the normalized weights. The training objective combines the binary cross-entropy loss between the ground truth labels $\mathbf{y}_{gt}$ and the weighted prompt probabilities $\hat{\mathbf{y}}$, with an entropy regularization term to encourage well-calibrated uncertainty:

$$\mathcal{L}_{\text{MLP}} = \text{BCE}(\hat{\mathbf{y}}, \mathbf{y}_{gt}) - \lambda \sum_{i=1}^{N} w_i \log(w_i) \tag{3}$$

During inference, the final predicted probabilities for each label are the weighted average of the individual prompt probabilities: $\hat{y} = \sum_{i=1}^{N} w_i p_i$. We leveraged the automatically labelled cases for the 8 selected labels to train the MLP model. We used 60% of the automatically labelled dataset, comprising 5794 cases, for training, 20% for validation and 20% for testing, ensuring this automated test split remained entirely distinct from the manually annotated expert evaluation set.

2.6 Evaluation Setup

We evaluated our approach on the manually annotated test dataset (462 reports, with 23 organ–finding combinations), excluding the 50 cases used to tune the linear weights. Model agreement was primarily quantified using Cohen's Kappa, a statistical metric that measures inter-rater agreement for categorical items while adjusting for the agreement occurring by chance alone. To establish a baseline, we simulated using any single prompt from our set of six, computing performance metrics for each prompt independently and then averaging the results.

We compared Uniform, Linear, and MLP-weighted ensembles, as well as our proposed Agent Decision Model. To further examine the effect of model capacity, we included an ablation variant ("Small Agent") in which the large Llama 3.1 model was replaced with the smaller Llama 3.1-8B model while keeping all other components identical. Each model was evaluated both before and after uncertainty filtering, with two filtering criteria: (1) excluding samples with uncertainty $\geq$0.5, and (2) limiting exclusions to at most 20% of cases if their uncertainty exceeded 0.5, accomplished by sorting the entire evaluation set by descending uncertainty scores and truncating the exclusion pool at the 20% rank boundary to maximize data retention.

3 Results

3.1 Uncertainty-aware prediction without filtering

Table 1 shows that aggregating multiple prompts significantly outperforms using a single prompt. The baseline (single-prompt) model has the lowest metrics, particularly F1=0.2699 and Kappa=0.1790. In contrast, the agent-based approach achieves the best overall balance, with F1=0.3967 and Kappa=0.3006, suggesting it handles ambiguity effectively. Although the MLP model attains the highest accuracy (0.8605) and precision (0.3772), its recall is lower (0.3977). Both uniform and linear methods also surpass the baseline, with linear showing slightly higher accuracy (0.8454) for similar F1, emphasizing the benefits of weight-optimized prompt ensembles. Notably, the Small Agent variant achieved the highest recall (0.7076), indicating strong sensitivity to positive findings. However, this came at the cost of reduced precision and overall agreement (Kappa=0.2402), suggesting it tends to over-predict positives compared with the other ensemble methods.

3.2 Uncertainty Histograms

Figures 2 and 3 show the uncertainty histograms for two representative labels from the four methods, with the Agent model providing the clearest sep-

Table 1: Average results across all labels before threshold.

Model	Accuracy	F1	Precision	Recall	Kappa
Baseline	0.8243	0.2699	0.3046	0.3097	0.1790
Uniform	0.8320	0.3873	0.3360	0.5336	0.2979
Linear	0.8454	0.3847	0.3463	0.4878	0.2988
MLP	**0.8605**	0.3643	**0.3772**	0.3977	0.2863
Agent	0.8022	**0.3967**	0.3131	0.6437	**0.3006**
Small Agent	0.7310	0.3555	0.2626	**0.7076**	0.2402

Table 2: Average median uncertainty values across methods.

Method	Median Uncertainty	
	Correct Predictions	Incorrect Predictions
Uniform	0.4081	0.6153
Linear	0.2668	0.4694
MLP	0.2261	0.5717
Agent	**0.0**	**0.5108**
Small Agent	0.0	0.5

aration between correct and incorrect predictions. Additional histograms appear in the supplementary materials.

Table 2 presents the average median uncertainty across all 23 labels, illustrating that a well-calibrated model should exhibit low uncertainty for correct predictions and high uncertainty for incorrect ones. The Agent model yields an average median uncertainty of 0.0 for correct predictions and about 0.5 for incorrect ones, indicating superior calibration compared to the other methods. While this strong separation highlights the model's effectiveness in distinguishing between reliable and unreliable predictions, the consistent 0.0 value for correct predictions also raises the possibility of discretization effects.

3.3 Uncertainty-aware prediction with filtering

Tables 3 and 4 present model performance after filtering out cases with uncertainty $\geq$0.5. Table 3 imposes no cap on excluded cases, while Table 4 limits exclusion to at most 20%. In both scenarios, removing high-uncertainty cases improves accuracy, F1, and Cohen's Kappa. Detailed per-label results can be found in the supplementary materials.

With no exclusion cap (Table 3), the MLP model achieves the highest accuracy (92.42%), whereas the Agent approach yields the best F1 (47.87

Under the 20% cap (Table 4), MLP again leads in accuracy (90.87%), and the Agent method retains the highest F1 (44.94%) and recall (72.95%), and Cohen's Kappa (0.3683). These results highlight the Agent method's strong emphasis on capturing true positives while maintaining reliable calibration, even with minimal case exclusions. The performance improvements achieved by filtering high-uncertainty cases demonstrate enhanced reliability but come at the cost of reduced report coverage, underscoring a fundamental trade-off between accuracy and scalability in automated workflows.

Table 3: Results after applying uncertainty threshold $\geq$ 0.5.

Model	Acc	F1	Prec	Recall	Kappa	Excluded Cases
Baseline	0.824	0.269	0.304	0.309	0.179	0
Uniform	0.918	0.463	0.445	0.532	0.420	200 (43.3%)
Linear	0.897	0.426	0.400	0.508	0.370	96 (20.9%)
MLP	**0.924**	0.403	**0.458**	0.407	0.364	120 (26.0%)
Agent	0.894	**0.478**	0.397	**0.661**	**0.425**	153 (33.1%)
Sm. Agent	0.851	0.421	0.331	0.685	0.353	185 (40.1%)

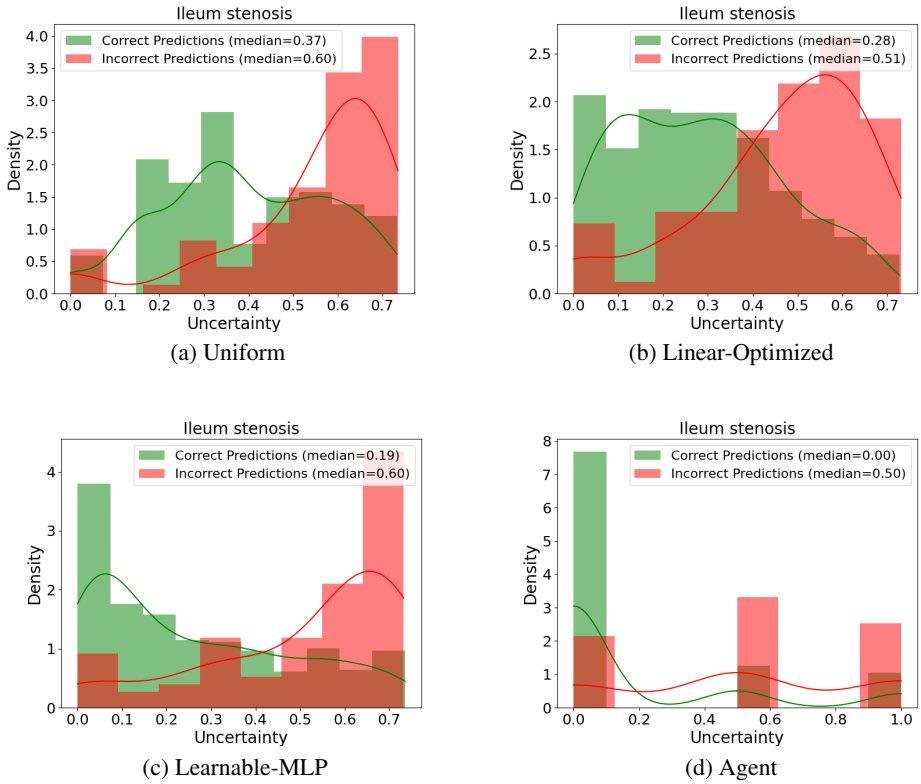
Table 4: Results after threshold $\geq$ 0.5 (max 20% exclusions).

Model	Acc	F1	Prec	Recall	Kappa	Excluded Cases
Baseline	0.824	0.269	0.304	0.309	0.179	0
Uniform	0.878	0.426	0.387	0.530	0.361	92 (19.9%)
Linear	0.890	0.212	0.393	0.507	0.361	80 (17.5%)
MLP	**0.908**	0.393	**0.449**	0.400	0.345	90 (19.5%)
Agent	0.828	**0.449**	0.352	**0.729**	**0.368**	92 (19.9%)
Sm. Agent	0.775	0.388	0.291	0.725	0.288	92 (19.9%)

3.4 Ablation Study: Effect of Agent Model Size

To assess the impact of the agent's language model size on performance, we conducted an ablation experiment replacing the large Llama3.1-70B model with a smaller Llama3.1-8B variant while keeping all other components identical. This setup allowed us to isolate the contribution of model capacity to decision quality and uncertainty calibration.

Across all metrics (Tables 1–4), the *Small Agent* consistently underperformed the full *Agent* model. Before uncertainty filtering (Table 1), the Small Agent achieved the highest recall (0.7076) but showed reduced accuracy and agreement (accuracy 0.7310; Cohen's Kappa 0.2402) relative to the full Agent (accuracy 0.8022; Cohen's Kappa 0.3006). Its higher sensitivity thus came at the cost of lower precision and F1, indicating a tendency toward over-detection. After uncertainty filtering ($u < 0.5$) and under the 20% exclusion cap (Tables 3–4), both models improved, yet the full Agent



(a) Uniform (b) Linear-Optimized

(c) Learnable-MLP (d) Agent

Figure 2: Uncertainty Histograms for Ileum stenosis Computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

retained superior accuracy, F1, and Cohen's Kappa (e.g., accuracy 0.8285 vs. 0.7751; F1 0.4494 vs. 0.3881; Cohen's Kappa 0.3683 vs. 0.2888). These results suggest larger LLM capacity yields better-calibrated, more balanced decisions, while smaller agents may favor sensitivity.

4 Discussion

In this study, we evaluated prompt-ensemble–based uncertainty awareness for structured data extraction from radiology reports using LLMs. Aggregating predictions from multiple prompts and filtering high-uncertainty cases outperformed single-prompt methods, notably boosting F1 and Cohen's Kappa. Among the tested approaches, the agent-based method provided the best overall balance in F1 and Cohen's Kappa by integrating multiple prompts and accounting for their consistency. Meanwhile, the MLP model attained the highest precision and accuracy, useful for scenarios demanding fewer false positives, but its lower recall indicates some missed positive cases. This trade-off underscores the importance of selecting models based on task needs.

Additionally, the agent approach demonstrated superior calibration, distinguishing correct and incorrect predictions more effectively than other methods. Entropy-based metrics further confirmed the agent's robust handling of complex prompt outputs, enhancing prediction reliability. Excluding high-uncertainty cases improved accuracy, F1, and Cohen's Kappa, showing the value of uncertainty-aware filtering for better alignment between model confidence and prediction correctness. Setting class-specific thresholds at the output layer can improve recall for minority classes while preserving precision for majority classes, thus optimizing performance for the target application.

While filtering high-uncertainty cases demonstrably improves performance metrics, this comes at the cost of reduced automated coverage, as a portion of reports are flagged for exclusion. This presents a critical trade-off for real-world clinical deployment: the system performs better on the cases it does process but offers no direct assistance for the cases it flags as uncertain. Clinicians would still need robust workflows to manually review these excluded reports, potentially impacting workflow efficiency—particularly in the most com-

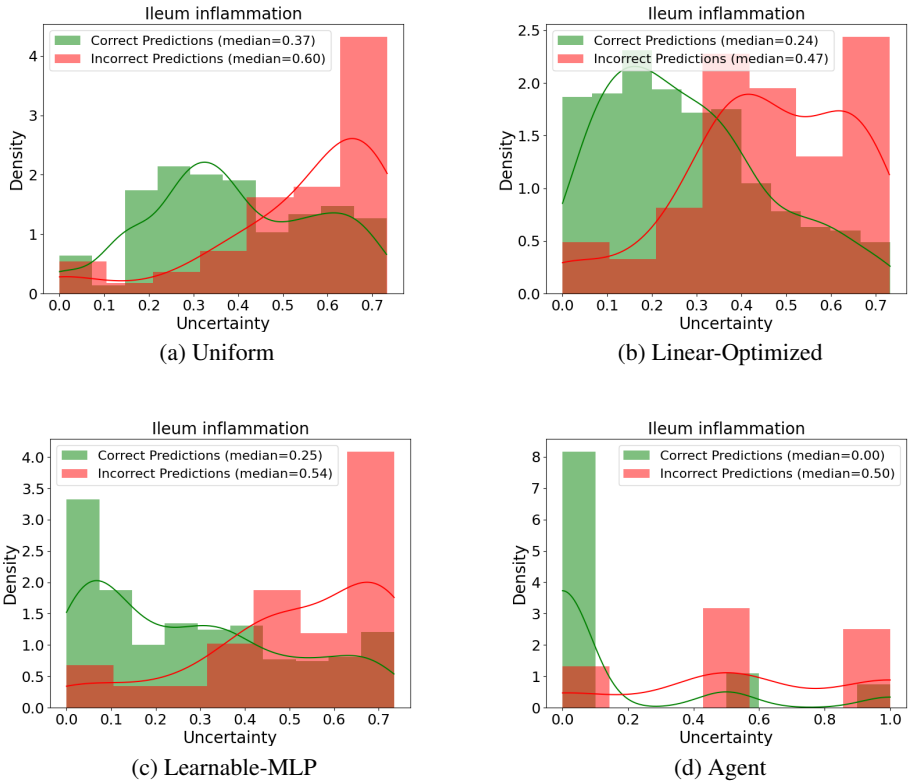


Figure 3: Uncertainty Histograms for Ileum inflammation Computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable- MLP Weights, and (d) Agent-based decision.

plex cases where automated support may be most needed. Implementing such a system in a clinical setting would necessitate defining clear workflows for handling cases flagged as uncertain. These reports could be automatically routed for mandatory manual review by a radiologist or technician, ensuring that no potentially critical information is missed.

Large language models (LLMs) efficiently automate structured data extraction from radiology reports by converting unstructured text into clinically relevant insights (Busch et al., 2024; Doshi et al., 2024; Reichenpfader et al., 2024). However, ensuring a trustworthy application through robust uncertainty quantification remains an open challenge. While Zeng et al. (Zeng et al., 2024) demonstrated the potential of multiple interacting LLM agents with specialized tasks to improve performance in medical applications, their role in quantifying uncertainty for reliable use remains underexplored. This study is the first to introduce a generalizable agent-based approach that quantifies uncertainty and fosters the trustworthy use of LLMs for structured data extraction from radiology reports.

While domain-specific medical LLMs, such as those fine-tuned on extensive biomedical text, hold significant promise for structured data extraction in healthcare, their applicability can be severely limited in low-resource languages like Hebrew. As noted, Llama 3.1 itself was not trained on Hebrew. However, even dedicated medical LLMs, if primarily trained on English or other well-represented languages, may struggle more profoundly than a multilingual base model when faced with Hebrew medical reports. Their domain-specific fine-tuning, while beneficial for known languages, can potentially make them less adaptable to languages outside their training corpus, exacerbating the performance gap due to the language barrier. This underscores the importance of methods like prompt ensembles and agent-based uncertainty awareness when deploying LLMs for medical text analysis in such linguistically challenging environments.

4.1 Limitations and Future Work

Despite promising results, several limitations merit discussion. First, while the focus was on radiology reports, extending these methods to LLM-based analysis of other medical reports, such as

pathology reports or electronic medical records, requires additional validation. Second, optimizing ensemble techniques for highly imbalanced datasets remains a challenge. Future clinical deployments will require targeted strategy modifications to address extreme class imbalances, such as class-weighted loss optimization for learnable models, threshold tuning specifically calibrated to maximize minority class sensitivity, or cost-sensitive agent prompting protocols. Future work should explore tailored approaches to address these limitations and investigate the application of uncertainty quantification in broader clinical scenarios. Another limitation of our comparison is the use of a significantly larger model (Llama 3 - 70B) for the agent compared to the base LLM (Llama 3.1 - 8B) and the models used in entropy-based methods. While the agent performs a different function (synthesizing outputs rather than initial extraction), its increased capacity could contribute to its superior performance metrics and calibration, representing a potentially unbalanced comparison. This size discrepancy might skew performance relative to simpler entropy-based methods, as the higher parameter count inherently grants stronger linguistic synthesis capabilities independent of the ensembling technique.

5 Conclusion

In conclusion, our findings demonstrate the effectiveness of prompt ensembles-based uncertainty awareness in enhancing LLM performance for structured data extraction in radiology. The agent-based approach emerged as particularly robust, achieving superior results. Incorporating uncertainty quantification not only improves reliability but also facilitates interpretability, paving the way for more impactful and trustworthy AI applications in healthcare.

Acknowledgments

The study was sponsored by the Leona M. and Harry B. Helmsley Charitable Trust. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

AI@Meta. 2024. Llama 3 Model Card.

David H Bruining, Ellen M Zimmermann, Edward V Loftus Jr, William J Sandborn, Cary G Sauer, Scott A Strong, and Society of Abdominal Radiology Crohn's Disease-Focused Panel. 2018. Consensus recommendations for evaluation, interpretation, and utilization of computed tomography and magnetic resonance enterography in patients with small bowel crohn's disease. *Radiology*, 286(3):776–799.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, and 1 others. 2024. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, pages 1–14.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rushabh Doshi, Kanhai S Amin, Pavan Khosla, Simar S Bajaj, Sophie Chheang, and Howard P Forman. 2024. Quantitative evaluation of large language models to streamline radiology report impressions: A multimodal retrospective analysis. *Radiology*, 310(3):e231593.

Mira Y Friedman, Maya Leventer-Roberts, Joseph Rosenblum, Nir Zigman, Iris Goren, Vered Mourad, Natan Lederman, Nurit Cohen, Eran Matz, Doron Z Dushnitzky, Nirit Borovsky, Moshe B Hoshen, Gili Focht, Malka Avitzour, Yael Shachar, Yehuda Chowers, Rami Eliakim, Shomron Ben-Horin, Shmuel Odes, and 7 others. 2018. Development and validation of novel algorithms to identify patients with inflammatory bowel diseases in israel: an epi-iirn group study. *Clinical Epidemiology*, pages 671–681.

Melissa A Haendel, Christopher G Chute, and Peter N Robinson. 2018. Classification, ontology, and precision medicine. *New England Journal of Medicine*, 379(15):1452–1462.

Saeed Hassanpour and Curtis P Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29–39.

L Hazan, G Focht, N Gavrielov, R Reichart, C Friss, R Cytter Kuint, D Turner, and M Freiman. 2024a. P269 harnessing natural language processing for structured information extraction from radiology reports in crohn's disease: A nationwide study from the epi-iirn. *Journal of Crohn's and Colitis*, 18(Supplement_1):i633–i634.

Liam Hazan, Gili Focht, Naama Gavrielov, Roi Reichart, Talar Hagopian, Mary-Louise C Greer, Ruth Cytter Kuint, Dan Turner, and Moti Freiman. 2024b. Leveraging prompt-learning for structured information extraction from crohn's disease radiology reports in a low-resource language. *arXiv preprint arXiv:2405.01682*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *IEEE Transactions on Software Engineering*, 51(2):413–429. ArXiv:2307.10236 [cs].

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6.

Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, Adam E Flanders, Matthew P Lungren, David S Mendelson, Jeffrey D Rudie, and 1 others. 2019. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781–791.

J Martijn Nobel, Koos van Geel, and Simon GF Robben. 2022. Structured reporting in radiology: a systematic review to explore its potential. *European radiology*, pages 1–18.

Sebastian Nowak, Benjamin Wulff, Yannik C Layer, Maike Theis, Alexander Isaak, Babak Salam, Wolfgang Block, Daniel Kuetting, Claus C Pieper, Julian A Luetkens, and 1 others. 2025. Privacy-ensuring open-weights large language models are competitive with closed-weights gpt-4o in extracting chest radiography findings from free-text reports. *Radiology*, 314(1):e240895.

OpenAI. 2024. ChatGPT (GPT-4).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155 [cs].

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *NPJ Digital Medicine*, 7(1):222.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities. *arXiv preprint arXiv:2407.07080*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *arXiv preprint*. ArXiv:2009.01325 [cs].

Hao Sun. 2024. Supervised Fine-Tuning as Inverse Reinforcement Learning. *arXiv preprint*. ArXiv:2403.12017 [cs].

Idit Tessler, Amit Wolfovitz, Eran E. Alon, Nir A. Gecel, Nir Livneh, Eyal Zimlichman, and Eyal Klang. 2024. ChatGPT's adherence to otolaryngology clinical practice guidelines. *European archives of oto-rhino-laryngology: official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, 281(7):3829–3834.

Francesco Tonolini, Jordan Massiah, Nikolaos Aletras, and Gabriella Kazai. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models.

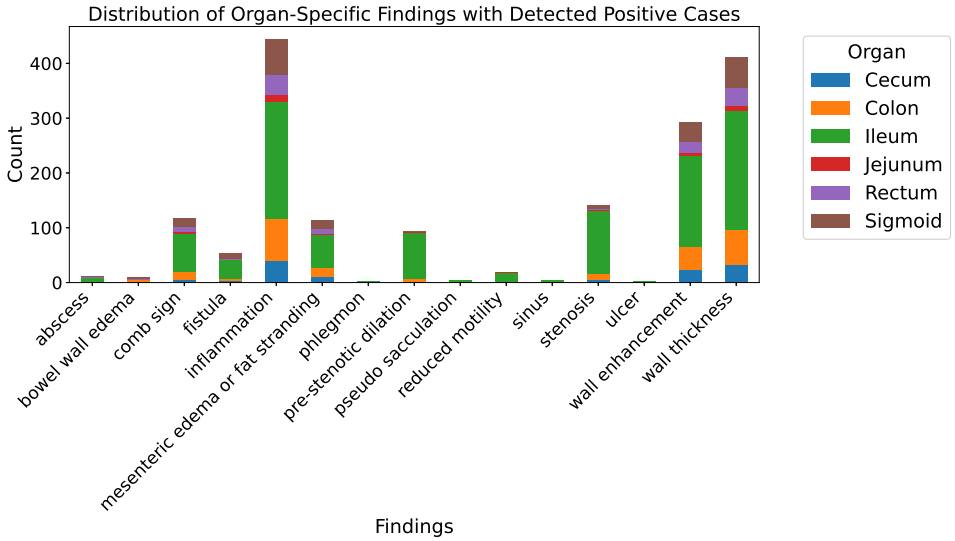
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. ArXiv:1706.03762 [cs].

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

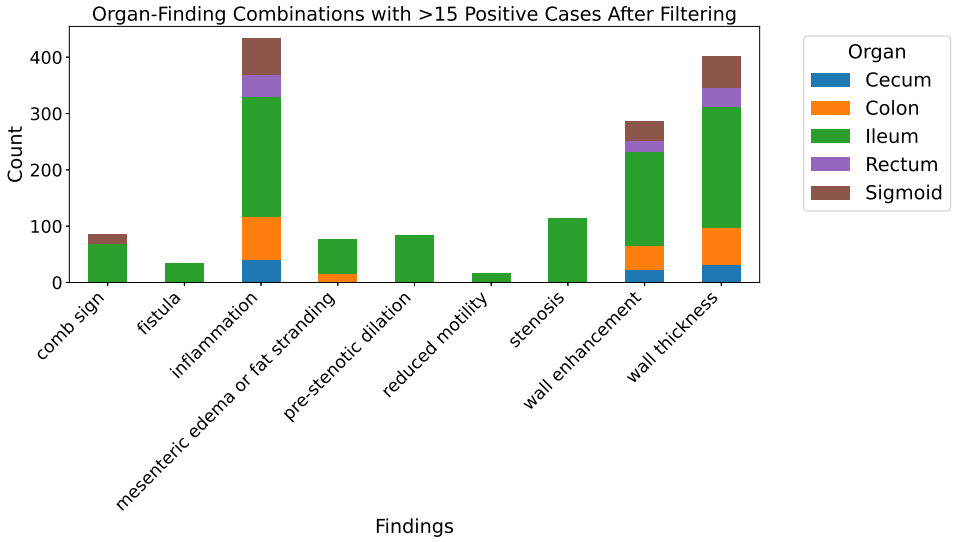
Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint*. ArXiv:2306.13063 [cs].

Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. 2024. Enhancing llms for impression generation in radiology reports through a multi-agent system. *arXiv preprint arXiv:2412.06828*.

A Data Distribution



(a) Organ-specific findings with positive cases.



(b) Combinations with >15 positive cases.

Figure 4: Distribution of annotated findings.

B SMP-BERT Performances

Table 5: SMP-BERT performances for the 23 selected labels on SZMC test-set.

Label	Accuracy	Precision	Recall	F1	Kappa	Roc-AUC
Cecum inflammation	0.9310	0.7729	0.6573	0.6959	0.3950	0.6573
Cecum wall enhancement	0.9482	0.6864	0.6864	0.6864	0.3729	0.6864
Cecum wall thickness	0.9568	0.8239	0.8611	0.8413	0.6827	0.8611
Colon inflammation	0.9051	0.8228	0.7739	0.7952	0.5910	0.7739
Colon mesenteric edema/fat	0.9568	0.4784	0.5000	0.4889	0.0000	0.5000
Colon wall enhancement	0.9310	0.8060	0.6905	0.7314	0.4654	0.6905
Colon wall thickness	0.9310	0.8939	0.8134	0.8468	0.6945	0.8134
Ileum comb sign	0.9482	0.9135	0.8722	0.8912	**0.7826 sequence**	0.8722
Ileum fistula	0.9396	0.7808	0.9164	0.8310	0.6639	0.9164
Ileum inflammation	0.8793	0.8745	0.8786	0.8763	**0.7527**	0.8786
Ileum mesenteric edema/fat	0.8965	0.8666	0.6512	0.6987	0.4101	0.6512
Ileum pre-stenotic dilation	0.9568	0.9211	0.9365	0.9286	**0.8572**	0.9365
Ileum reduced motility	0.9827	0.9912	0.7500	0.8289	0.6588	0.7500
Ileum stenosis	0.9224	0.8883	0.9137	0.8999	**0.8000**	0.9137
Ileum wall enhancement	0.9568	0.9514	0.9559	0.9535	**0.9071**	0.9559
Ileum wall thickness	0.9396	0.9416	0.9444	0.9396	**0.8796**	0.9444
Rectum inflammation	0.9396	0.8482	0.6619	0.7147	0.4345	0.6619
Rectum wall enhancement	0.9741	0.9867	0.7500	0.8266	0.6547	0.7500
Rectum wall thickness	0.9827	0.9909	0.8571	0.9121	**0.8245**	0.8571
Sigmoid comb sign	0.9913	0.9956	0.8333	0.8977	**0.7957**	0.8333
Sigmoid inflammation	0.8793	0.7483	0.6937	0.7156	0.4329	0.6937
Sigmoid wall enhancement	0.9310	0.7729	0.6573	0.6959	0.3950	0.6573
Sigmoid wall thickness	0.9482	0.9143	0.7868	0.8358	0.6729	0.7868

C Prompts and Agent Examples

Base LLM Extraction Prompts

1. Does the following radiology report indicate that the patient has {*label*}? Here is the report: {}. Answer shortly, format your answer as a JSON file using the following schema: {"Answer": "Yes/No", "Explanation": "str"}. Important: Only return a single piece of valid JSON text.
2. Does the radiology report below suggest that the patient is suffering from {*label*}? Here is the report: {}. Please provide a concise answer, formatted as a JSON object using the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
3. Can the presence of {*label*} be confirmed from the following radiology report? Here is the report: {}. Provide a brief response, formatted as a JSON object according to the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
4. Does the patient appear to have {*label*} based on the analysis of the radiology report provided? Here is the report: {}. Please answer succinctly, formatted as a JSON object in the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
5. Considering the radiology report presented, is there an indication that the patient has {*label*}? Here is the report: {}. Provide a brief answer, formatted as a JSON object according to the schema: {"Answer": "Yes/No", "Explanation": "str"}.
6. Is it possible to conclude from the following radiology report that the patient has {*label*}? Here is the report: {}. Please provide a concise response, formatted as a JSON object using the schema: {"Answer": "Yes/No", "Explanation": "str"}.

Agent Decision Model Example

Input (Multiple Answers from Base LLM):

```
- Answer: "No". Explanation: "The report does not mention Cecum inflammation,
but rather mentions air-filled bubbles in the soft tissue of the right abdomen,
likely originating from the small intestine."
- Answer: "Yes". Explanation: "The report mentions the presence of bubbles or
air pockets in the soft tissue of the right abdomen, which may be indicative
of inflammation in the cecum."
```

Agent Output (Llama 3-70B synthesis):

```
"Decision": "Likely No",
"Explanation":  "The majority of the answers suggest that there is no
inflammation in the Cecum, and the explanations provided are clear and
consistent.  The one answer that suggests yes is not strongly supported and
seems to be an over-interpretation of the report."
```

D Detailed Results Before Filtering

Table 6: Comparison of different methods **before filtering** using uncertainty threshold

Label	Prev.	F1 Score [%]					Accuracy [%]					Kappa [%]				
		Base	Uni	Lin	MLP	Agnt	Base	Uni	Lin	MLP	Agnt	Base	Uni	Lin	MLP	Agnt
Cecum inflammation	8.2%	25.4	28.1	29.5	28.8	33.0	83.1	79.0	81.3	82.9	80.7	0.16	0.18	0.20	0.20	0.24
Cecum wall enhance.	4.5%	9.5	12.0	12.0	11.4	11.4	84.7	80.9	84.1	86.5	76.6	0.03	0.05	0.05	0.05	0.04
Cecum wall thick.	6.4%	24.9	32.7	31.8	32.9	27.4	85.4	83.1	83.3	88.5	80.5	0.18	0.25	0.24	0.27	0.19
Colon inflammation	15.1%	37.0	44.9	44.1	42.7	43.4	77.6	74.0	75.3	75.1	69.0	0.23	0.30	0.30	0.28	0.27
Colon edema/fat	3.2%	4.8	8.8	10.0	13.7	18.6	91.5	91.1	92.2	94.5	86.7	0.01	0.04	0.06	0.11	0.14
Colon wall enhance.	8.2%	16.3	23.9	24.3	20.2	28.7	84.6	84.8	86.5	88.0	79.6	0.09	0.15	0.17	0.13	0.19
Colon wall thick.	12.7%	29.5	41.9	39.6	38.6	45.8	84.1	86.7	86.7	88.3	84.6	0.21	0.34	0.32	0.33	0.37
Ileum comb sign	14.5%	28.2	41.8	39.2	31.7	42.9	77.8	77.7	79.8	81.3	74.6	0.16	0.29	0.27	0.21	0.29
Ileum fistula	6.2%	33.7	48.4	51.2	54.2	43.1	87.8	89.3	91.7	93.0	87.4	0.28	0.43	0.47	0.50	0.37
Ileum inflammation	41.7%	60.5	86.1	84.1	84.6	86.9	73.8	88.5	87.2	87.8	88.5	0.43	0.76	0.73	0.74	0.76
Ileum edema/fat	11.9%	17.6	34.0	26.9	26.5	36.8	84.4	85.7	85.9	86.7	84.4	0.10	0.26	0.19	0.20	0.28
Ileum pre-stenotic	16.0%	27.0	40.8	42.1	40.6	39.5	73.0	68.6	71.4	77.9	62.3	0.12	0.23	0.26	0.27	0.20
Ileum reduced motil.	2.8%	17.1	26.3	33.3	28.0	20.0	87.5	87.8	92.2	92.2	82.6	0.13	0.22	0.30	0.24	0.15
Ileum stenosis	22.0%	39.5	56.5	53.6	53.6	64.6	78.5	80.7	80.5	82.0	82.2	0.27	0.44	0.41	0.42	0.53
Ileum wall enhance.	31.8%	40.0	65.8	70.4	50.0	67.6	68.8	76.7	78.3	73.1	75.7	0.21	0.48	0.53	0.32	0.48
Ileum wall thick.	42.8%	56.8	79.6	77.5	73.4	88.3	72.3	84.6	83.3	81.3	90.0	0.39	0.67	0.64	0.60	0.79
Rectum inflammation	7.1%	21.3	27.2	28.2	31.4	26.0	85.2	82.6	85.7	86.7	77.9	0.14	0.19	0.21	0.24	0.17
Rectum wall enhance.	3.6%	8.9	20.6	18.8	11.7	20.6	89.9	90.0	90.6	93.5	85.0	0.05	0.16	0.14	0.08	0.15
Rectum wall thick.	5.8%	25.0	33.8	29.7	37.7	36.8	89.6	89.8	88.7	92.8	85.9	0.20	0.28	0.24	0.33	0.30
Sigmoid comb sign	3.4%	16.0	21.9	25.0	18.8	18.3	88.3	87.6	90.9	90.6	82.6	0.11	0.17	0.21	0.14	0.13
Sigmoid inflamm.	12.5%	29.6	40.0	37.2	39.3	36.0	75.4	71.4	72.2	75.3	63.8	0.16	0.26	0.23	0.26	0.20
Sigmoid enhance.	7.1%	17.9	27.3	23.3	23.7	33.8	84.9	85.0	87.2	90.2	80.5	0.11	0.19	0.16	0.18	0.25
Sigmoid wall thick.	11.0%	33.0	47.7	52.2	43.3	41.7	86.8	87.2	88.5	89.8	83.1	0.26	0.40	0.45	0.38	0.32

E Detailed Results After Filtering (0.5)

Table 7: Comparison of different methods after filtering using uncertainty threshold $<$ **0.5**

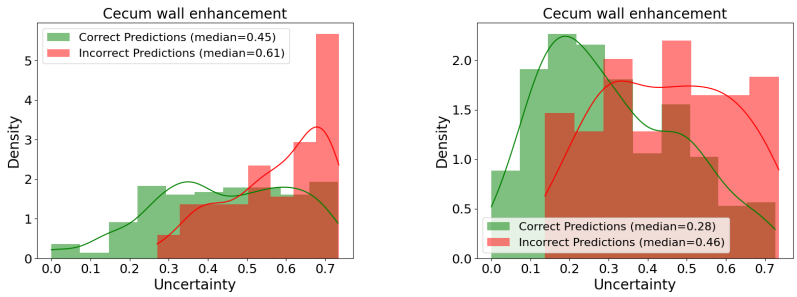
Label	F1 Score					Accuracy					Kappa					# Cases Kept				
	Base	Uni	Lin	MLP	Agnt	Base	Uni	Lin	MLP	Agnt	Base	Uni	Lin	MLP	Agnt	Base	Uni	Lin	MLP	Agnt
Cecum inflammation	25.4	37.2	28.5	35.2	33.3	83.1	90.2	88.4	93.0	88.9	0.16	0.32	0.23	0.31	0.29	462	276	346	317	327
Cecum wall enhance.	9.5	14.2	15.3	7.1	13.6	84.7	90.1	87.9	92.1	88.0	0.03	0.10	0.10	0.03	0.08	462	243	366	331	317
Cecum wall thick.	24.9	59.2	47.6	41.6	40.0	85.4	95.9	93.1	95.7	91.4	0.18	0.57	0.44	0.39	0.36	462	273	323	333	315
Colon inflammation	37.0	54.2	49.6	51.7	51.7	77.6	81.3	80.7	83.6	77.3	0.23	0.43	0.39	0.42	0.39	462	289	369	343	314
Colon edema/fat	4.8	20.0	18.1	26.6	28.5	91.5	97.4	95.5	97.1	95.2	0.01	0.18	0.15	0.25	0.26	462	308	402	387	317
Colon wall enhance.	16.3	14.8	22.2	16.2	17.0	84.6	91.5	89.2	91.5	87.1	0.09	0.10	0.16	0.12	0.11	462	271	390	365	304
Colon wall thick.	29.5	53.3	46.6	39.0	58.1	84.1	94.9	91.5	93.0	92.5	0.21	0.50	0.42	0.36	0.54	462	276	379	361	309
Ileum comb sign	28.2	50.0	33.8	44.8	45.0	77.8	91.1	86.6	91.5	84.3	0.16	0.45	0.26	0.40	0.36	462	180	351	321	249
Ileum fistula	33.7	72.7	52.3	68.9	71.7	87.8	96.7	95.0	97.6	96.4	0.28	0.71	0.50	0.67	0.70	462	279	402	384	309
Ileum inflammation	60.5	93.1	87.7	89.2	96.3	73.8	94.5	90.8	92.7	97.0	0.43	0.88	0.80	0.83	0.93	462	296	381	347	344
Ileum edema/fat	17.6	16.6	27.4	0.0	33.3	84.4	93.1	90.3	92.8	92.4	0.10	0.14	0.22	-0.01	0.29	462	293	383	375	317
Ileum pre-stenotic	27.0	60.6	55.2	55.0	47.7	73.0	84.0	81.5	87.3	75.9	0.12	0.51	0.44	0.48	0.35	462	163	298	246	245
Ileum reduced motil.	17.1	40.0	32.0	33.3	44.4	87.5	96.2	95.5	97.6	94.6	0.13	0.38	0.30	0.32	0.42	462	237	382	339	278
Ileum stenosis	39.5	69.2	62.0	57.9	79.2	78.5	91.3	87.9	91.2	92.9	0.27	0.64	0.54	0.53	0.75	462	278	364	332	314
Ileum wall enhance.	40.0	69.5	73.5	57.1	67.1	68.8	85.6	83.7	86.1	82.1	0.21	0.60	0.62	0.48	0.55	461	195	252	281	275
Ileum wall thickness	56.8	91.9	83.8	81.6	95.5	72.3	95.4	90.3	91.6	96.9	0.39	0.88	0.77	0.76	0.93	462	244	353	323	329
Rectum inflammation	21.3	36.8	32.8	38.0	36.9	85.2	91.5	88.3	92.6	87.5	0.14	0.32	0.26	0.34	0.31	462	284	419	352	329
Rectum wall enhance.	8.9	0.0	20.6	0.0	28.5	89.9	94.8	94.1	96.2	93.5	0.05	-0.02	0.17	-0.01	0.26	462	273	393	378	309
Rectum wall thickness	25.0	58.8	40.0	38.0	50.0	89.6	97.5	94.1	96.3	93.9	0.20	0.57	0.37	0.36	0.47	462	288	360	357	328
Sigmoid comb sign	16.0	18.7	20.5	23.0	32.4	88.3	90.7	92.5	94.5	92.4	0.11	0.14	0.17	0.20	0.29	462	282	414	368	331
Sigmoid inflamm.	29.6	42.2	43.5	43.4	45.7	75.4	78.1	79.2	82.2	77.2	0.16	0.31	0.33	0.34	0.35	462	238	338	293	281
Sigmoid enhance.	17.9	33.3	31.5	28.5	37.3	84.9	90.2	92.9	93.2	81.8	0.11	0.28	0.27	0.25	0.29	462	258	351	359	310
Sigmoid wall thick.	33.0	60.6	50.8	47.3	52.1	86.8	95.6	92.3	94.4	90.5	0.26	0.58	0.46	0.45	0.47	462	299	381	363	350

F Detailed Results After Filtering (Max 20%)

Table 8: Comparison of different methods after filtering up to **20%** of data, using uncertainty threshold < 0.5

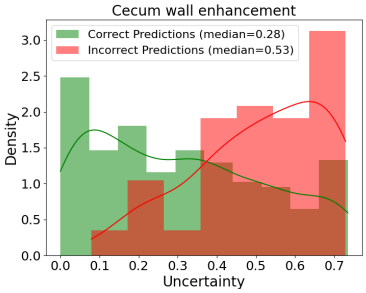
Label	F1 Score					Accuracy					Kappa					# Cases Kept				
	Base	**Uni**	**Lin**	**MLP**	**Agnt**	**Base**	**Uni**	**Lin**	**MLP**	**Agnt**	**Base**	**Uni**	**Lin**	**MLP**	**Agnt**	**Base**	**Uni**	**Lin**	**MLP**	**Agnt**
Cecum inflammation	25.4	31.3	33.3	35.0	36.3	83.1	84.5	88.1	90.0	82.9	0.16	0.24	0.27	0.29	0.29	462	370	370	370	370
Cecum wall enhance.	9.5	15.6	15.3	15.0	14.8	84.7	85.4	88.1	90.8	78.3	0.03	0.09	0.10	0.10	0.08	462	370	370	370	370
Cecum wall thick.	24.9	38.7	42.6	35.8	30.9	85.4	89.7	90.5	93.2	81.8	0.18	0.34	0.37	0.32	0.24	462	370	370	370	370
Colon inflammation	37.0	49.0	49.2	50.0	50.2	77.6	77.5	80.5	81.6	73.7	0.23	0.36	0.38	0.39	0.36	462	370	370	370	370
Colon edema/fat	4.8	16.6	18.1	26.6	17.8	91.5	94.5	95.5	97.1	87.5	0.01	0.14	0.15	0.25	0.13	462	370	402	387	370
Colon wall enhance.	16.3	28.0	22.2	16.2	30.1	84.6	88.9	89.2	91.6	80.0	0.09	0.22	0.16	0.12	0.21	462	370	390	370	370
Colon wall thick.	29.5	47.7	46.6	40.9	54.9	84.1	90.5	91.5	92.9	87.5	0.21	0.42	0.42	0.38	0.48	462	370	379	370	370
Ileum comb sign	28.2	41.5	35.4	36.6	47.8	77.8	84.0	86.2	87.8	77.0	0.16	0.32	0.27	0.30	0.35	462	370	370	370	370
Ileum fistula	33.7	65.3	52.3	68.9	53.0	87.8	95.1	95.0	97.6	89.4	0.28	0.62	0.50	0.67	0.48	462	370	402	384	370
Ileum inflammation	60.5	90.6	87.7	88.2	92.6	73.8	92.4	90.8	91.8	93.7	0.43	0.84	0.80	0.82	0.87	462	370	381	370	370
Ileum edema/fat	17.6	29.7	27.4	0.0	45.9	84.4	91.0	90.3	92.8	87.2	0.10	0.25	0.22	-0.01	0.38	462	370	383	375	370
Ileum pre-stenotic	27.0	46.1	46.7	46.6	43.9	73.0	73.5	75.9	82.7	62.7	0.12	0.31	0.33	0.36	0.24	462	370	370	370	370
Ileum reduced motil.	17.1	31.5	32.0	37.5	21.6	87.5	92.9	95.5	97.2	82.4	0.13	0.28	0.30	0.36	0.17	462	370	382	370	370
Ileum stenosis	39.5	63.7	61.6	60.6	76.4	78.5	86.7	87.5	89.4	89.1	0.27	0.55	0.54	0.54	0.69	462	370	370	370	370
Ileum wall enhance.	40.0	66.9	72.7	53.9	74.1	68.8	79.9	80.7	79.6	79.9	0.21	0.52	0.58	0.41	0.58	461	369	369	369	369
Ileum wall thickness	56.8	86.8	83.7	79.2	93.8	72.3	91.0	90.0	88.1	95.1	0.39	0.80	0.76	0.71	0.89	462	370	370	370	370
Rectum inflammation	21.3	38.0	32.8	36.0	29.1	85.2	89.4	88.3	91.3	81.6	0.14	0.32	0.26	0.31	0.21	462	370	419	370	370
Rectum wall enhance.	8.9	7.6	20.6	0.0	23.1	89.9	93.5	94.1	96.2	85.6	0.05	0.04	0.17	-0.01	0.19	462	370	393	378	370
Rectum wall thickness	25.0	42.8	41.0	36.3	44.7	89.6	93.5	93.7	96.2	87.2	0.20	0.39	0.37	0.34	0.39	462	370	370	370	370
Sigmoid comb sign	16.0	18.1	20.5	23.0	23.5	88.3	90.2	92.5	94.5	85.9	0.11	0.14	0.17	0.20	0.18	462	370	414	370	370
Sigmoid inflamm.	29.6	39.7	43.6	40.3	41.4	75.4	73.7	78.3	79.1	65.6	0.16	0.27	0.33	0.30	0.25	462	370	370	370	370
Sigmoid enhance.	17.9	33.3	31.5	28.5	37.3	84.9	90.2	92.9	93.2	81.8	0.11	0.28	0.27	0.25	0.29	462	370	370	370	370
Sigmoid wall thick.	33.0	51.7	50.8	48.7	49.4	86.8	92.4	92.3	94.3	88.3	0.26	0.47	0.46	0.46	0.43	462	370	381	370	370

G Uncertainty Histograms

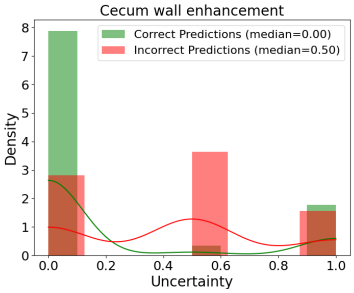


(a) Uniform

(b) Linear-Optimized

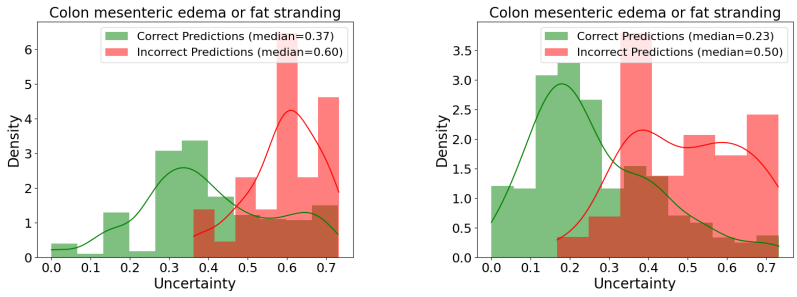


(c) Learnable-MLP



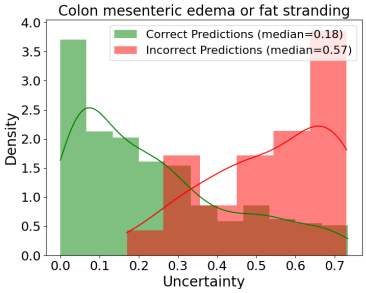
(d) Agent

Figure 5: Uncertainty Histograms for Cecum Wall Enhancement computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

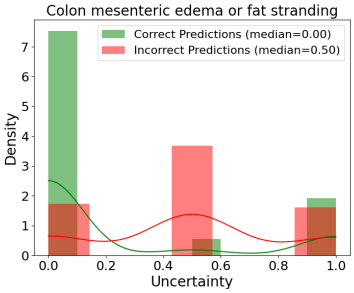


(a) Uniform

(b) Linear-Optimized

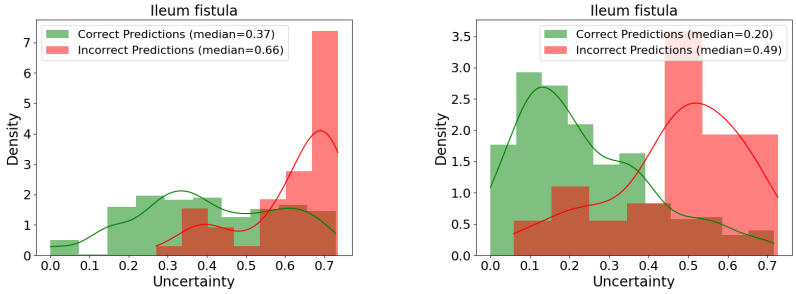


(c) Learnable-MLP



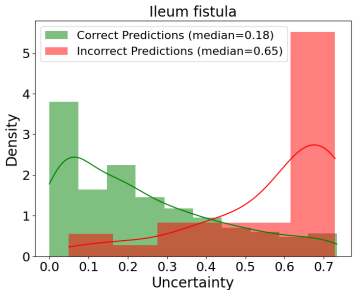
(d) Agent

Figure 6: Uncertainty Histograms for Colon Mesenteric Edema or Fat Stranding computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

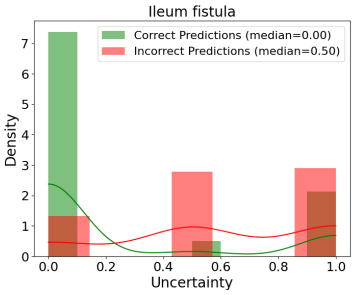


(a) Uniform

(b) Linear-Optimized

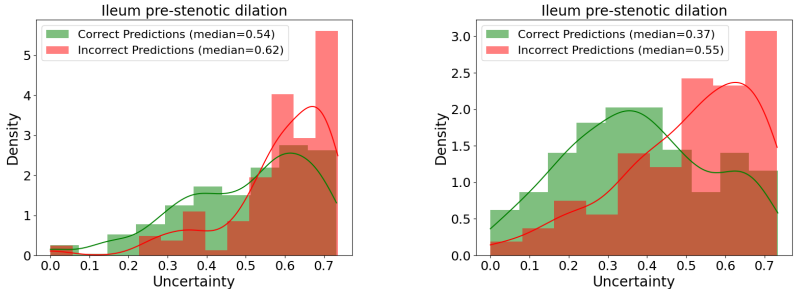


(c) Learnable-MLP



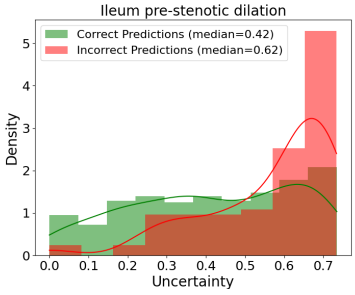
(d) Agent

Figure 7: Uncertainty Histograms for Ileum Fistula computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

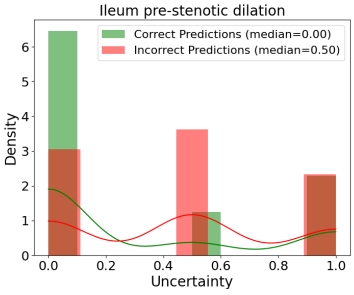


(a) Uniform

(b) Linear-Optimized

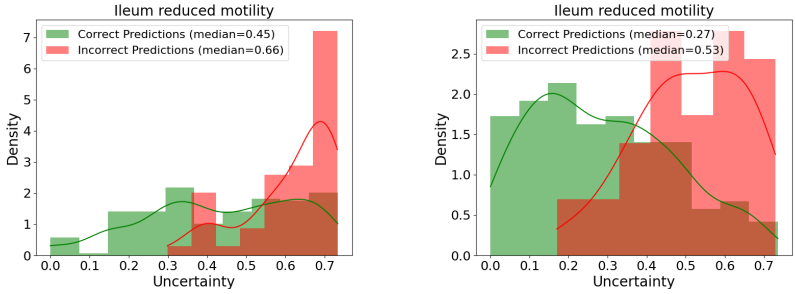


(c) Learnable-MLP



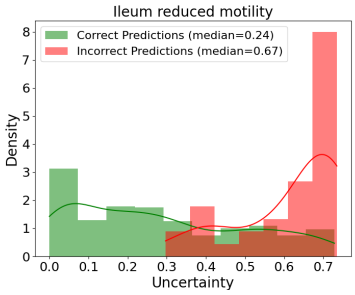
(d) Agent

Figure 8: Uncertainty Histograms for Ileum Pre-Stenotic Dilation computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

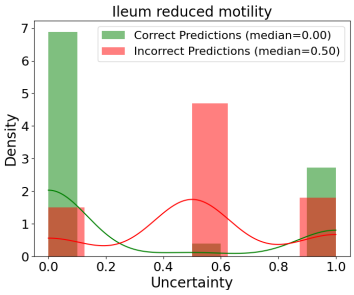


(a) Uniform

(b) Linear-Optimized

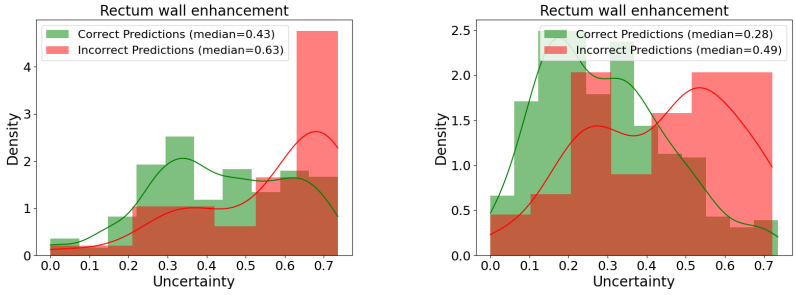


(c) Learnable-MLP



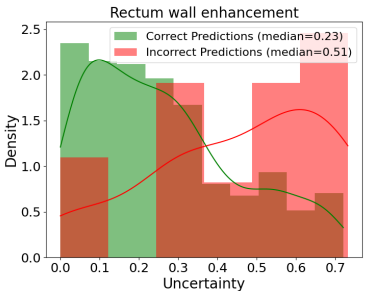
(d) Agent

Figure 9: Uncertainty Histograms for Ileum Reduced Motility computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

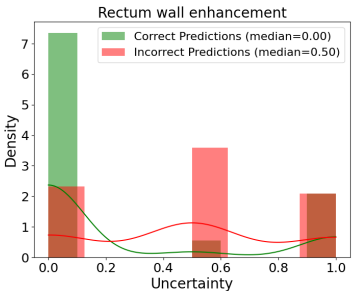


(a) Uniform

(b) Linear-Optimized

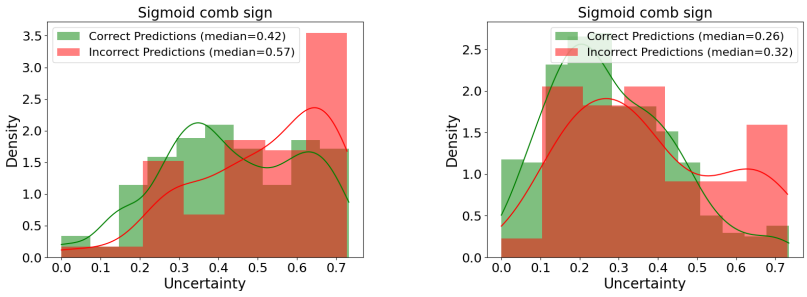


(c) Learnable-MLP

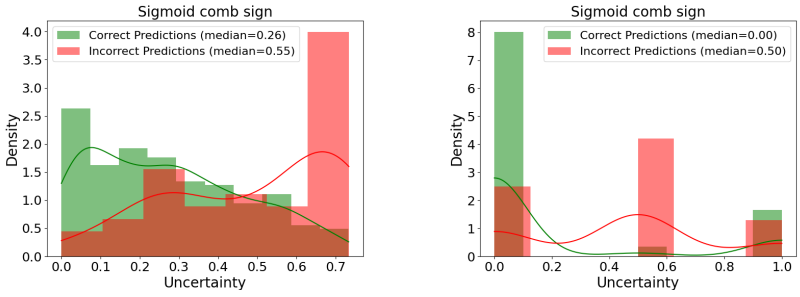


(d) Agent

Figure 10: Uncertainty Histograms for Rectum Wall Enhancement computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



(a) Uniform (b) Linear-Optimized



(c) Learnable-MLP (d) Agent

Figure 11: Uncertainty Histograms for Sigmoid Comb-Sign computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

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

Ken Yano and Hiroya Takamura
National Institute of Advanced Industrial Science and Technology
Artificial Intelligence Research Center
{yano.ken, takamura.hiroya}@aist.go.jp

Abstract

NER requires token-level classification using both left and right context, which makes encoder-only models like BERT naturally well-suited for the task. Decoder-only LLMs, by contrast, use causal masking during training, so their token representations lack right-side context, limiting their effectiveness on structured prediction tasks like NER despite their strong general capabilities. To address this, the authors propose fine-tuning decoder-only LLMs with causal attention replaced by full attention, combined with label-supervised discriminative training. While similar ideas exist in prior work, those studies were limited in scope. This work evaluates seven LLMs across four model families (Gemma, Qwen2.5, Llama3.1, Llama3.2) and compares full fine-tuning against LoRA. Results show that the proposed approach with an appropriate LoRA configuration outperforms encoder baselines (BERT, RoBERTa, DeBERTa), and achieves strong NER performance without auxiliary data or architectural modifications, though it does not reach SOTA on BC5CDR and CoNLL2003.

1 Introduction

Named Entity Recognition (NER) in the biomedical domain is a foundational task for extracting structured knowledge from clinical and scientific text, enabling downstream applications such as drug-disease relationship extraction, clinical trial mining, and pharmacovigilance. Biomedical NER is traditionally treated as a sequence-labeling task using a tagging scheme such as BIO, enabling detection and classification of entity spans — such as diseases, chemicals, genes, and clinical procedures — via token classification. It is therefore crucial to correctly classify each token using the context before and after it, a particular challenge in biomedical text where entity boundaries are often ambiguous, and terminology is highly specialized.

Encoder-only models such as BioBERT, PubMedBERT, and DeBERTa are trained to predict the masked token given the surrounding context, making their learned token representations well-suited to token classification tasks such as biomedical NER. On the other hand, decoder-only LLMs are trained to generate text in an autoregressive manner using a causal mask. As a result, the learned token representation contains no information about the context to the right of each token, which is a structural disadvantage for span-level extraction tasks.

Nonetheless, decoder-only LLMs trained on massive amounts of text — including biomedical literature, clinical notes, and scientific corpora — can solve various tasks using an instruction prompt, with or without in-context examples, and demonstrate strong capabilities across a wide range of biomedical understanding tasks. However, this generative solution method has not shown sufficient strength for structure prediction tasks such as NER in both general and biomedical domains. In particular, in the biomedical domain, it requires precise extraction of entity spans — including nested and discontinuous mentions extraction common in clinical text — and accurate classification of fine-grained biomedical entity types such as chemical compounds, disease mentions, and genomic variants.

To address these issues, we propose a label-supervised discriminative LLM fine-tuning method that replaces causal attention with a full-attention mask during both training and inference, adapting decoder-only LLMs for NER. Similar methods have been proposed in previous work (Li et al., 2023; Dukić and Šnajder, 2024); however, they examined only a limited number of LLM model families of a few different sizes. Our method also partially overlaps with LLM2Vec (BehnamGhader et al., 2024), where LLMs were effectively transformed into encoders. Our NER method cannot

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 312–325
July 3-4, 2026 ©2026 Association for Computational Linguistics

handle nested and discontinuous entity mentions, which are common in the biomedical domain and require span-based approaches (Tan et al., 2020; Su et al., 2022; Wang et al., 2022), which are beyond the scope of this work.

In this work, we evaluated seven decoder LLMs with varying sizes (from 0.5B to 8B) from four model families — Gemma, Qwen2.5, Llama3.1, and Llama3.2 — on general and biomedical NER benchmarks. For the baseline encoders, we evaluated BERT, RoBERTa, and DeBERTa. We also examined how two different training approaches, full fine-tuning and LoRA, affect the performance of the proposed method in both general and biomedical settings, where labeled data are often scarce and costly to annotate. Our results show that label-supervised discriminative LLM fine-tuning with full self-attention outperforms all baseline encoders when fine-tuned with an appropriate LoRA configuration. Although the results of our proposed method did not reach state-of-the-art on two NER benchmarks, BC5CDR and CoNLL2003, our models still demonstrate strong performance even when fine-tuned solely on each benchmark's training set, without relying on auxiliary corpora, domain-adaptive pretraining, or additional model architecture modifications.

2 Related Work

2.1 Named Entity Recognition

NER is traditionally tackled as a sequence labeling task, and numerous methods have been proposed (Lample et al., 2016; Ma and Hovy, 2016). With the emergence of large-scale transformer-based pretrained models, fine-tuning has become the standard for achieving better performance (Peters et al., 2018; Devlin et al., 2019).

The major approaches using transformer-based models can be categorized as follows: (1) supervised fine-tuning using encoder-based models such as BERT (Devlin et al., 2019) and its variants (Liu et al., 2019), (2) supervised fine-tuning using encoder-decoder models such as T5 (Wang et al., 2023; Paolini et al., 2021), (3) few-shots in-context learning using LLMs (Ding et al., 2021; Monajatipoor et al., 2024; Wang et al., 2025), and (4) supervised fine-tuning using LLMs (Xu et al., 2025; Dukić and Šnajder, 2024; Li et al., 2023).

Another line of effort focuses on the structure prediction task for longer contexts, such as multi-document relation extraction (Tan et al., 2022; Xue et al., 2024; Chen et al., 2020; Amalvy et al., 2023).

2.2 Generative methods by LLMs

Even though decoder-only LLMs have achieved SOTA performances on a variety of NLP tasks, their performance on NER still falls short of the supervised encoder baselines. This is because NER is a sequence-labeling task, whereas decoder-only LLMs are text-generation models.

Although several LLM-based NERs have been proposed, they still perform subpar with supervised encoder models. For example, GPT-NER (Wang et al., 2025) claims to perform on par with the encoder model. However, their method has some limitations, such as the ability to extract only one entity type at a time and the need for many in-context examples to match the performance of supervised encoder models.

There are also issues with converting structured information, such as NER, to a serialized form generated by LLMs. Without guardrails, parsing errors are a significant concern.

To overcome these issues, several approaches exist for generating structured output from LLMs, including prompt engineering, constrained decoding, grammar-based generation, and tool-augmented generation. Even though structured output from LLMs helps mitigate parsing errors, it significantly impacts inference time.

2.3 Discriminative methods by LLMs

This paper is closely related to the work by Li et al. (2023) and Dukić and Šnajder (2024), where the causal attention of LLMs is replaced by full attention to solve structure prediction tasks. Dukić and Šnajder (2024) proposed selectively removing a causal attention mask by identifying some of the decoder layers, instead of all decoder layers, to yield the best performance. However, this search space is enormous if the number of decoder layers n is large.

To reduce the search space to a manageable size, Dukić and Šnajder (2024) proposed dividing the entire set of decoder layers into four layer groups, each with eight consecutive decoder layers, where n is 32. This leaves them $2^4 = 16$ possible search patterns to evaluate.

However, it remains unclear whether this consecutive grouping of an equal number of decoder layers is optimal for achieving the best NER performance. Moreover, their ablation studies did

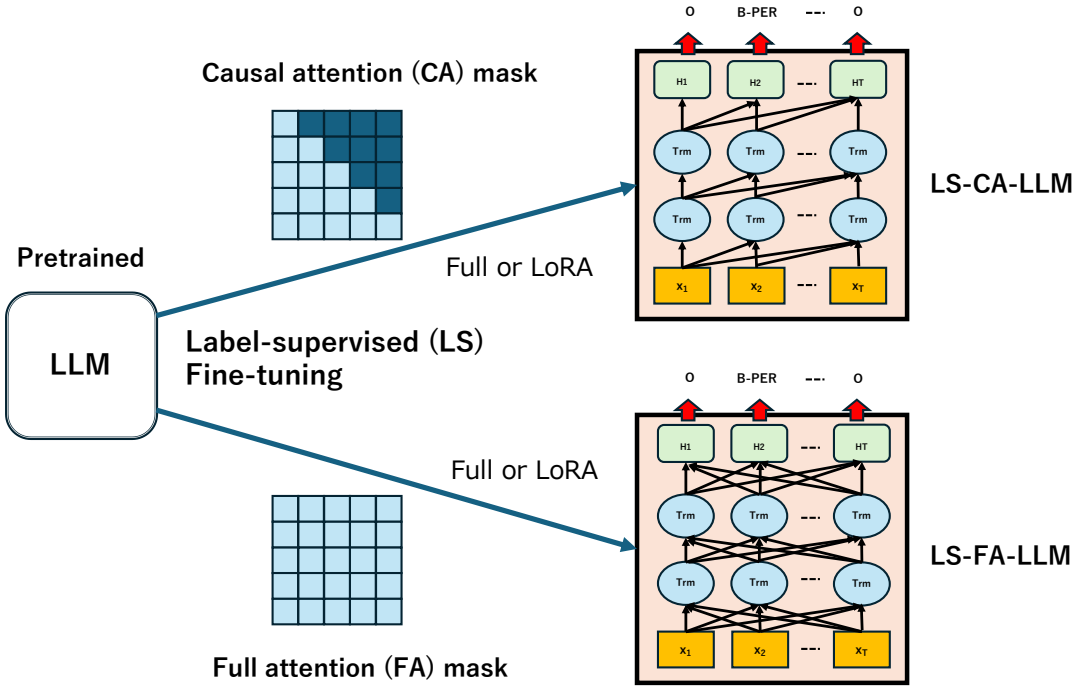


Figure 1: Label supervised (LS) LLM fine-tuning for NER task. The upper right figure shows the default LS fine-tuning with a causal attention mask. The lower-right figure shows the proposed LS fine-tuning, which replaces causal attention with a full attention mask across all decoder blocks. We denote the LS fine-tuned models using causal attention (CA) and full attention (FA) masks as LS-CA-LLM and LS-FA-LLM, respectively.

not identify consistent replacement patterns for decoder layers that yielded significantly better results across different decoder models and NER tasks. Hence, we replaced the causal attention mask with a full attention mask across all decoder layers.

3 Token Classification by Decoder-only LLM

We use LLMs not as generative models but as token classifiers to solve NER tasks, similar to encoder models, in which NER is performed by classifying each token into predefined BIO tags.

Token classification LLM models are defined by replacing the language modeling head with a token classification head initialized with pretrained LLM weights. The token classification head is initialized to random values before fine-tuning.

During fine-tuning, token classification LLMs are trained to classify each token using supervised labels. Here, the self-attention of original LLMs is defined using a causal attention mask, so the learned token representation contains no information about the context to the right of each token. This is a disadvantage, as both past and future token information is essential for classifying each token's label.

So our proposed method replaces the causal attention (CA) mask with a full attention (FA) mask across all decoder transformer blocks, ensuring that learned token representations capture context not only from past but also from future tokens, as shown in Fig. 1. We define label supervised (LS) fine-tuning of LLMs using CA and FA mask by LS-CA-LLM and LS-FA-LLM, respectively, and, for a specific LLM, we append **LS-CA-** and **LS-FA-** as prefixes to the name of LLMs in the following discussion to differentiate them.

Implementation of LS-FA-LLMs

We use HuggingFace Transformers to implement the proposed NER models. For a particular LLM model family, the corresponding token classification model is defined separately in the library. For instance, for Llama, the corresponding token classification model is defined as *LlamaForTokenClassification* in the library.

To replace the causal attention of LLMs, originally implemented for their corresponding token classifier models, such as *LlamaForTokenClassification*, with a full attention mask, we manually modified the source code. Currently, the source modification is ad hoc, meaning we need to manu-

ally change the source code. We leave, making the change systematic for future work.

4 Experimental Setup

4.1 NER datasets

We used two NER datasets for our experiments. One is BC5CDR (Wei et al., 2016) and the other is CoNLL2003 (Tjong Kim Sang and De Meulder, 2003). BC5CDR defines two entity types, chemical and disease, and CoNLL2003 defines four entity types, person, organization, location, and miscellaneous. Details of each dataset are provided in Table 1.

We also constrained our experimental conditions to avoid using any auxiliary datasets that would boost LLM performance. Hence, we train each model using only the corresponding training and validation datasets and evaluate its performance on the corresponding testing dataset.

4.2 Examined baseline encoders and decoder-only LLMs

For the baselines, we used the following encoder models in our experiments: BERT (bert-base-cased) (Devlin et al., 2019), RoBERTa (roberta-base) (Liu et al., 2019), and DeBERTa (deberta-v3-large) (He et al., 2021).

For the decoder models, we selected four LLM model families: Gemma, Qwen2.5, Llama3.2, and Llama3.1. For each model family, we selected one or two models with different model sizes.

Specifically, we evaluated seven LLMs across various model sizes. Gemma-2B and Gemma-7B (Team et al., 2024), Qwen2.5-0.5B and Qwen2.5-7B, Llama3.2-1B and Llama3.2-3B and Llama3.1-8B (Grattafiori et al., 2024).

State-of-the-art models

For BC5CDR, we evaluated BioBERT_v1.1 (Lee et al., 2019), which is specifically fine-tuned for the biomedical domain, on the same experimental setting as other models. In addition, we compare OpenMED's SOTA results (Panahi, 2025) for reference.

For CoNLL2003, we compare the SOTA results of ACE (Automated Concatenation of Embeddings, proposed by Wang et al. (Wang et al., 2021), fine-tuned on CoNLL2003, for reference.

4.3 Training of label-supervised sequence labeling

LS-FA-LLMs, LS-CA-LLMs, and the baseline encoder models were trained for sequence labeling by adding a token classification head on top of the last hidden layer. The token classification head, initialized to random values, was trained alongside the underlying model parameters during fine-tuning.

In our experiments, all models are trained without parameter-efficient fine-tuning (PEFT) technique and with LoRA (Hu et al., 2021). By these experiments, we can analyze the results in four ways: either CA or FA attention mask with either full or LoRA training.

By using LoRA, we can reduce the trainable parameters of LS-CA-LLMs and LS-FA-LLMs to a much smaller size, even compared with the full parameter sizes of encoder models, as shown in Table 4.

Moreover, using LoRA, we were able to fit all LS-CA-LLMs and LS-FA-LLMs during fine-tuning with a consistent device batch size of 8 on a single H200 GPU with 141 GB of memory. To fully train LS-CA-LLMs and LS-FA-LLMs over 7B, we use multiple H200 GPUs to avoid out-of-memory errors. The details of training parameters and LoRA configuration are in Appendix A.

We used the same default LoRA configuration for all training. In the later result section (5.2), we investigate how different LoRA configurations affect the results. We conducted additional experiments with different LoRA configurations for LS-FA-Llama3.2-3B and LS-FA-Llama3.1-8B, which performed better among other models.

4.4 Comparison with generative NER by instruction fine-tuned LLMs

To verify the effectiveness of our proposed method, we also evaluate NER performance using the same LLMs with instruction-fine-tuning (IFT) to solve the same NER tasks in a generative manner.

To fine-tune IFT LLMs, we used the same LoRA configuration and hyperparameters as those used to train LS-FA-LLMs. However, to prevent overfitting to the training dataset, we use early_stopping and stop training when the last 10 consecutive updates do not improve the evaluation loss.

For NER instruction, we adopt a format similar to Alpaca (Taori et al., 2023). The detailed instruction samples are provided in Appendix B. We use **IFT-** as the prefix to denote instruction-fine-tuned

Dataset	Train	Validation	Test	Entity Types
BC5CDR	5,228	5,330	5,865	Chemical, Disease
CoNLL2003	14,041	3,250	3,453	PER,ORG,LOC,MISC

Table 1: The statistics for the number of samples for BC5CDR (Wei et al., 2016) and CoNLL2003 (Tjong Kim Sang and De Meulder, 2003). PER, ORG, LOC, MISC specify person, organization, location, and miscellaneous, respectively.

LLMs such as IFT-Gemma-2B in the results.

5 Results

We first present the NER results by comparing baseline encoders with LS-CA-LLMs and LS-FA-LLMs, all of which are trained with full or default LoRA settings. Then we show performance variability across different LoRA configurations.

In the last section, we show comparative results between generative methods by IFT-LLMs and LS-FA-LLMS with baselines.

5.1 Comparison of baseline encoders and LS-CA-LLMs and LS-FA-LLMs

Table 2 shows precision, recall, and test micro F1 scores of BC5CDR by baseline encoders, LS-CA-LLMs, and LS-FA-LLMs. At the top, we show the SOTA results from OpenMed for reference, along with the results of domain-adapted BioBERT_v1.1. To evaluate sequence labeling results, we used seqeval. The first three columns are results of full training, and the last three columns are results of LoRA training. For each column, the best score is indicated in **bold** font. The scores with a yellow background specify the best score across full and LoRA training. In both full and LoRA training, LS-FA-LLM models significantly outperform the corresponding LS-CA-LLM models by 5-15 points, indicating that the causal attention mask is the main cause of low performance on NER tasks.

When trained with LoRA, LS-CA-LLMs and LS-FA-LLMs consistently outperform fully trained models, indicating that LoRA training is highly effective for fine-tuning LLMs with relatively small training datasets such as BC5CDR.

On the other hand, we cannot confirm a similar performance gain of LoRA over full training for baseline encoders, especially for BERT and RoBERTa. We hypothesize that LoRA training is effective only when the training dataset is not large enough relative to the model's capacity, in which case full training may lead to overfitting.

We confirm that the results of LS-FA-Llama3.2-3B and LS-FA-Llama3.1-8B, trained with LoRA, outperform all baseline encoders and the domain-adapted BioBERT_v1.1, excluding SOTA f1 results by OpenMed (Panahi, 2025).

Similarly, Table 3 shows precision, recall, and test micro F1 scores of CoNLL2003 by baseline encoders, LS-CA-LLMs, and LS-FA-LLMs. At the top, we present the SOTA f1 result by ACE+Fine-tune (Wang et al., 2021) for reference.

We see similar trends to those seen in the previous BC5CDR results in Table 2. Excluding SOTA f1 by ACE+Fine-tune, LS-FA-Llama3.2-3B and LS-FA-Llama3.1-8B, trained with LoRA, outperform all the baseline encoders in all metrics.

5.2 Variability of performance on different LoRA configurations

We evaluate how different LoRA configurations affect performance. We used LS-FA-Llama3.2-3B and LS-FA-Llama3.1-8B, which outperformed the baselines and other LS-FA-LLM results. Table 4 shows the results of five different LoRA configurations for these two models. The row in lightgrey is the result of LoRA configuration used in Tables 2 and 3.

In the table, the columns of "r", "alpha", "dropout", and "modules" specify the LoRA configuration items: attention dimension (rank), LoRA scaling, dropout probability, and target modules, respectively. Column "#trainable params" specified the number of trainable params, and its percentage of the model size. By adopting LoRA, the number of trainable parameters in LS-FA-LLMs is much lower than in encoder models trained without LoRA.

The results of these two LS-FA-LLMs vary across different LoRA configurations. However, we confirm that LoRA configurations with relatively large $rank \geq 32$ and $alpha = 2 \times rank$ targeted for all linear modules achieve higher performance.

Training	Full			LoRA		
BC5CDR	Precision	Recall	F1	Precision	Recall	F1
SOTA						
OpenMed (Panahi, 2025)	F1 = 0.961 (Chemical), F1 = 0.912 (Disease)					
BioBERT_v1.1 (Lee et al., 2019)	0.874	**0.918**	**0.895**	0.860	0.908	0.884
Baseline Encoders						
BERT	0.845	0.872	0.858	0.828	0.875	0.851
RoBERTa	0.871	0.890	0.880	0.842	0.889	0.865
DeBERTa	**0.876**	0.889	0.882	0.877	0.909	0.893
LS-CA-LLM						
LS-CA-Gemma-2B	0.716	0.701	0.708	0.747	0.776	0.761
LS-CA-Gemma-7B	0.622	0.628	0.625	0.768	0.801	0.784
LS-CA-Qwen2.5-0.5B	0.628	0.612	0.620	0.666	0.681	0.674
LS-CA-Qwen2.5-7B	0.669	0.696	0.682	0.710	0.730	0.720
LS-CA-Llama3.2-1B	0.630	0.602	0.616	0.692	0.704	0.698
LS-CA-Llama3.2-3B	0.634	0.621	0.627	0.710	0.732	0.721
LS-CA-Llama3.1-8B	0.660	0.655	0.657	0.719	0.739	0.729
LS-FA-LLM						
LS-FA-Gemma-2B	0.848	0.853	0.851	0.873	0.899	0.886
LS-FA-Gemma-7B	0.737	0.730	0.734	0.876	0.904	0.890
LS-FA-Qwen2.5-0.5B	0.849	0.825	0.837	0.858	0.868	0.863
LS-FA-Qwen2.5-7B	0.765	0.735	0.750	0.875	0.902	0.888
LS-FA-Llama3.2-1B	0.799	0.799	0.799	0.876	0.901	0.889
LS-FA-Llama3.2-3B	0.843	0.839	0.841	**0.892**	0.906	0.899
LS-FA-Llama3.1-8B	0.810	0.798	0.804	0.887	**0.920**	**0.904**

Table 2: BC5CDR: Precision, recall and micro F1 scores for the baseline encoders, LS-CA-LLMs and LS-FA-LLMs. At the top, we include SOTA results by OpenMed (Panahi, 2025) for reference and BioBERT_v1.1 (Lee et al., 2019). The first three columns are results of full training, and the last three columns are results of LoRA training. For each column, the best score is indicated in **bold** font. The scores with a yellow background specify the best score across full and LoRA training.

5.3 Comparison with instruction fine-tuned LLMs approach

Table 5 shows the comparison with instruction fine-tuned (IFT) counterparts with label-supervised full-attention LLMs and baseline encoders. We also add zero-shot results by OpenAI GPT-4.1 as a baseline (see in Appendix C for details). The best scores are indicated in bold font.

To train LLMs using IFT, we adopt an Alpaca-based instruction format (Taori et al., 2023) for fine-tuning, as described in the Appendix B. Since the named entities are extracted from the output text, the performance of IFT is evaluated using set-based precision, recall, and F1 scores between the estimated and true sets of entities, ignoring the spans of the extracted entities. We use the exact match when comparing estimated and true results.

To compare with span-based results by label-supervised LLMs and encoders, we convert them to set-based metrics by aggregating extracted entities.

To train IFT models, we used the same LoRA configuration as the label-supervised full-attention counterparts.

First, IFT-LLM results outperformed GPT-4.1 in zero-shot settings, indicating that correctly extracting entities remains challenging for high-performance closed-weight LLMs. Larger IFT-LLM results on BC5CDR show competitive performance relative to baseline encoders; however, they fall short of the baseline encoders on CoNLL2003.

Results on LS-FA-LLMs show superior performance on both BC5CDR and CoNLL2003, outperforming both IFT counterparts and baseline encoders. These results, combined with the results on Tables 2 and 3, underscore the superiority of the proposed LS-FA-LLMs.

6 Analysis

6.1 Training and Inferencing time

Table 6 shows a comparison of training time among all models for one epoch using a single H200 GPU. The column 'rel' indicates the relative time based on BERT's time. In these experiments, baseline encoder models are trained without LoRA, and LS-FA-LLMs are trained with LoRA. We confirm that, even with LoRA, much longer training times are required for LS-FA-LLMs. The training time proportionally increases with the size of the original LLMs. When fine-tuning on 7B LS-FA-LLMs, about $16-18\times$ and $15-16\times$ training times are re-

Training	Full			LoRA		
CoNLL2003	Precision	Recall	F1	Precision	Recall	F1
SOTA						
ACE+Fine-tune (Wang et al., 2021)			F1 = 0.946			
Baseline Encoders						
BERT	0.901	0.915	0.908	0.903	0.918	0.910
RoBERTa	**0.908**	**0.924**	**0.915**	0.912	0.928	0.920
DeBERTa	0.904	0.917	0.911	0.903	0.918	0.910
LS-CA-LLM						
LS-CA-Gemma-2B	0.729	0.780	0.754	0.762	0.814	0.787
LS-CA-Gemma-7B	0.680	0.739	0.708	0.766	0.815	0.790
LS-CA-Qwen2.5-0.5B	0.609	0.667	0.636	0.656	0.714	0.684
LS-CA-Qwen2.5-7B	0.656	0.714	0.684	0.696	0.751	0.722
LS-CA-Llama3.2-1B	0.626	0.678	0.651	0.684	0.740	0.711
LS-CA-Llama3.2-3B	0.618	0.675	0.645	0.700	0.757	0.727
LS-CA-Llama3.1-8B	0.643	0.700	0.671	0.699	0.754	0.725
LS-FA-LLM						
LS-FA-Gemma-2B	0.873	0.883	0.878	0.914	0.929	0.922
LS-FA-Gemma-7B	0.778	0.817	0.797	0.922	0.932	0.927
LS-FA-Qwen2.5-0.5B	0.806	0.818	0.812	0.887	0.894	0.891
LS-FA-Qwen2.5-7B	0.851	0.861	0.856	0.910	0.929	0.920
LS-FA-Llama3.2-1B	0.851	0.871	0.861	0.915	0.927	0.920
LS-FA-Llama3.2-3B	0.845	0.859	0.852	**0.930**	**0.941**	**0.935**
LS-FA-Llama3.1-8B	0.773	0.795	0.784	0.922	0.936	0.929

Table 3: CoNLL2003: Precision, recall and micro F1 scores for the baseline encoders, LS-CA-LLMs and LS-FA-LLMs. At the top, we include SOTA results by ACE+Fine-tune (Wang et al., 2021) for reference. The first three columns are results of full training, and the last three columns are results of LoRA training. For each column, the best score is indicated in **bold** font. The scores with a yellow background specify the best score across full and LoRA training.

quired for BC5CDR and CoNLL2003, respectively, compared with BERT.

These results show a negative effect of using larger LLMs for NER, despite achieving higher performance.

However, if multiple GPUs are available, the training time of LS-FA-LLMs can be further reduced using a distributed parallel training method

Table 7 shows the mean and the standard deviation of inference time in milliseconds to process one sample among 10 randomly selected test samples contained in BC5CDR and CoNLL2003 using a single H200 GPU. Larger LS-FA-LLMs such as LS-FA-Gemma-7B and LS-FA-Llama3.1-8B take $\sim$ 10$\times$ inference time compared with BERT, and smaller LS-FA-LLMs proportionally take less inference time as their parameter size gets smaller. LS-FA-Qwen2.5-0.5B shows exceptionally longer inference time compared with models with similar sizes, such as LS-FA-Gemma-2B and LS-FA-Llama3.2-1B. We have yet to analyze this phenomenon and leave it for future work.

We note again that if multiple GPUs are available, the inference time of LS-FA-LLMs can be further reduced using a distributed parallel decoding method.

6.2 Error Analysis

Tables 11 and 12 show the error analysis of DeBERTa, LS-FA-Llama3.1-8B, and LS-CA-Llama3.1-8B for BC5CDR and CoNLL2003, respectively. NER error was classified into three classes: SPAN, TYPE, and DETECTION errors, and each error class was further classified into more detailed error types.

In both analyses, we confirmed that the counts of errors for each error type were lower in LS-FA-Llama3.1-8B than in DeBERTa. This indicates that LS-FA-Llama3.1-8B has acquired more robust named-entity recognition knowledge than DeBERTa. LS-CA-Llama3.1-8B shows poor performance in both analyses, confirming that the causal attention mechanism is the bottleneck for NER.

In the analysis of LS-FA-Llama3.1-8B and DeBERTa, TYPE error counts are much lower in BC5CDR than in CoNLL2003, indicating greater ambiguity among entity types in CoNLL2003.

The above results and analysis confirm that the decoder LLMs' weakness in structured precision tasks, such as NER, can be mitigated by our proposed method.

Model	BC5CDR F1	CoNLL2003 F1	r	alpha	dropout %	modules	#trainable params
LS-FA-Llama3.2-3B	0.883	0.914	8	16	0.05	q,k,v,o	4.60M (0.143%)
	0.889	0.918	16	32	0.05	q,k,v,o	9.19M (0.285%)
	0.899	**0.935**	32	64	0.05	all linear	48.60M (1.490%)
	0.898	0.928	64	128	0.05	all linear	97.30M (2.940%)
	0.900	0.927	128	256	0	all linear	195.00M (5.710%)
LS-FA-Llama3.1-8B	0.888	0.919	8	16	0.05	q,k,v,o	6.84M (0.091%)
	0.893	0.922	16	32	0.05	q,k,v,o	13.70M (0.182%)
	0.904	0.929	32	64	0.05	all linear	83.90M (1.110%)
	0.904	0.929	64	128	0.05	all linear	168.00M (2.190%)
	0.900	**0.932**	128	256	0	all linear	336.00M (4.280%)

Table 4: Test micro F1 scores for LS-FA-Llama3.2-3B and LS-FA-Llama3.1-8B with five different LoRA configurations, including the one used for previous experiments shown in light gray background. Columns "r" and "alpha" specify rank and alpha values of LoRA parameters. The "modules" specifies the targeted modules and "all linear" specifies all linear modules: q_proj,k_proj,v_proj,o_proj,down_proj,gate_proj and up_proj.

7 Conclusion

We address weaknesses in decoder LLMs for structure prediction tasks such as NER by focusing on the attention mechanism and adopting a discriminative approach with supervised labels. Specifically, we demonstrate that this weakness can be mitigated by replacing causal attention with a full-attention mask during both training and inference.

However, our results show that this change to the self-attention mechanism alone does not yield superior NER performance compared with strong baselines such as RoBERTa and DeBERTa. Specifically, we verified that instead of full fine-tuning, using LoRA with a suitable configuration is key to boosting the performance. This is especially true for large LLMs because full fine-tuning on a relatively small training dataset, such as BC5CDR or CoNLL2003, can easily lead to overfitting, hindering performance on the test set.

Our results also suggested that linguistic knowledge acquired during pre-training of LLMs with causal attention is transferable even after the attention mechanism is switched to full-attention during fine-tuning. In our experiments, we used only the training subsets of the NER benchmarks for fine-tuning, so additional training with an auxiliary dataset could increase the performance even more.

Although our proposed best model outperformed all baseline encoders on BC5CDR and CoNLL2003, it requires much longer training and inference times than the baselines. Nonetheless, we verified that our proposed method is highly effective and can serve as an alternative encoder model to BERT-like models in downstream applications.

Since the LoRA configuration significantly affects the performance of the proposed methods, future work should investigate automated methods to obtain an optimal LoRA configuration.

Acknowledgments

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain".

Limitations

In our experiments, we conducted a single fine-tuning step for each model, using a fixed learning rate and other hyperparameters. With more computing resources, multiple runs with different training parameters might yield more reliable results.

Our experiments were conducted on two NER datasets, one in the biomedical domain and the other in the general domain, so further experiments on datasets in other domains and languages could yield additional insights.

References

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. The role of global and local context in named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 714–722, Toronto, Canada. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Yubo Chen, Chuhan Wu, Tao Qi, Zhigang Yuan, and Yongfeng Huang. 2020. Named entity recognition

		BC5CDR			CoNLL2003		
Model	**Training**	**Precision (set)**	**Recall (set)**	**F1 (set)**	**Precision (set)**	**Recall (set)**	**F1 (set)**
Baseline Encoders							
BERT	Full	0.861	0.877	0.869	0.914	0.914	0.914
RoBERTa	Full	0.882	0.896	0.889	0.917	0.921	0.919
DeBERTa	Full	0.887	0.896	0.892	0.915	0.916	0.915
IFT-LLM							
GPT-4.1	Zero-shot	0.670	0.505	0.576	0.388	0.669	0.491
IFT-Gemma-2B	LoRA	0.870	0.858	0.864	0.870	0.859	0.864
IFT-Gemma-7B	LoRA	0.878	0.876	0.877	0.876	0.873	0.874
IFT-Qwen2.5-0.5B	LoRA	0.803	0.804	0.804	0.807	0.808	0.808
IFT-Qwen2.5-7B	LoRA	0.881	0.878	0.880	0.882	0.874	0.878
IFT-Llama3.2-1B	LoRA	0.866	0.861	0.864	0.865	0.860	0.862
IFT-Llama3.2-3B	LoRA	0.879	0.862	0.870	0.879	0.861	0.869
IFT-Llama3.1-8B	LoRA	0.901	0.887	0.894	0.902	0.888	0.895
LS-FA-LLM							
LS-FA-Gemma-2B	LoRA	0.889	0.900	0.894	0.929	0.930	0.930
LS-FA-Gemma-7B	LoRA	0.887	0.906	0.896	0.932	0.933	0.933
LS-FA-Qwen2.5-0.5B	LoRA	0.874	0.874	0.874	0.899	0.897	0.898
LS-FA-Qwen2.5-7B	LoRA	0.893	0.902	0.898	0.920	0.925	0.923
LS-FA-Llama3.2-1B	LoRA	0.893	0.902	0.898	0.928	0.928	0.928
LS-FA-Llama3.2-3B	LoRA	**0.907**	0.909	0.908	**0.937**	**0.940**	**0.938**
LS-FA-Llama3.1-8B	LoRA	0.904	**0.920**	**0.912**	0.933	0.936	0.935

Table 5: Comparison with instruction fine-tuned (IFT) counterparts with label supervised full-attention LLMs and baseline encoders. Performance of IFT is evaluated by set-based precision, recall, and f1 between estimated entities and true entities using exact match. Token-based results by label-supervised LLMs and encoders are converted to set-based metrics. Best scores are in bold face.

	BC5CDR		CoNLL2003	
	[mm:ss]	rel.	[mm:ss]	rel.
BERT	00:39	1.0	01:08	1.0
RoBERTa	00:36	0.9	01:09	1.0
DeBERTa	01:47	2.8	03:27	3.1
LS-FA-Gemma-2B	03:30	5.4	05:49	5.2
LS-FA-Gemma-7B	11:47	18.3	18:29	16.4
LS-FA-Qwen2.5-0.5B	02:08	3.3	03:46	3.3
LS-FA-Qwen2.5-7B	11:25	17.8	17:21	15.4
LS-FA-Llama3.2-1B	02:13	3.5	03:58	3.5
LS-FA-Llama3.2-3B	05:20	8.3	08:23	7.4
LS-FA-Llama3.1-8B	11:48	18.4	17:54	15.9

Table 6: Elapsed time to train encoders and LS-FA-LLMs for BC5CDR and CoNLL2003, each for one epoch, using one H200 GPU. The column 'rel' indicates relative time based on BERT's time. Encoder models are trained without LoRA and LS-FA-LLMs are trained with LoRA.

	BC5CDR	CoNLL2003
	Time [msec]	Time [msec]
BERT	2.15 ± 0.02	2.82 ± 0.10
RoBERTa	3.03 ± 0.08	3.54 ± 0.03
DeBERTa	17.88 ± 0.46	18.23 ± 0.34
LS-FA-Gemma-2B	8.29 ± 0.08	7.91 ± 0.06
LS-FA-Gemma-7B	23.93 ± 0.59	26.48 ± 0.37
LS-FA-Qwen2.5-0.5B	15.55 ± 3.77	11.32 ± 0.26
LS-FA-Qwen2.5-7B	21.31 ± 0.31	24.42 ± 0.68
LS-FA-Llama3.2-1B	6.23 ± 0.21	12.85 ± 0.29
LS-FA-Llama3.2-3B	13.20 ± 0.20	11.57 ± 0.69
LS-FA-Llama3.1-8B	23.99 ± 0.45	26.39 ± 0.81

Table 7: Elapsed inference time in milliseconds to process an example out of 10 randomly selected test samples for BC5CDR and CoNLL2003

in multi-level contexts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 181–190, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of

decoder-only llms for sequence labeling. *Preprint*, arXiv:2401.14556.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *Preprint*, arXiv:2310.01208.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *Preprint*, arXiv:2404.07376.

Maziyar Panahi. 2025. Openmed ner: Open-source, domain-adapted state-of-the-art transformers for biomedical ner across 12 public datasets. *Preprint*, arXiv:2508.01630.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *ArXiv*, abs/2101.05779.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Preprint*, arXiv:1802.05365.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *Preprint*, arXiv:2208.03054.

Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. Boundary enhanced neural span classification for nested named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9016–9023.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, Jun Huang, and Ming Gao. 2022. SpanProto: A two-stage span-based prototypical network for few-shot named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3476, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2643–2660, Online. Association for Computational Linguistics.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database: The Journal of Biological Databases and Curation*, 2016.

Weilu Xu, Renfei Dang, and Shujian Huang. 2025. LLM's weakness in NER doesn't stop it from enhancing a stronger SLM. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 170–175, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. AutoRE: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220, Bangkok, Thailand. Association for Computational Linguistics.

A Training parameters

Baseline encoders, label-supervised LLMs, and instruction fine-tuned LLMs are trained using the same hyperparameters shown in Table 8. Tables 9 and 10 show default LoRA configurations used for our experiments for LLMs and encoders.

Hyper-parameter	Value
per_device_batch_size	8
gradient_accumulation_steps	1
num_train_epochs	10
learning_rate	5E-05
weight_decay	0.001
warmup_ratio	0
lr_scheduler_type	cosine
gradient_checkpointing	False
model_max_length	512
bf16	false

Table 8: Training parameters

LoRA configuration item	Value
lora_r	32
lora_alpha	64
lora_dropout	0.05
lora_bias	none
lora_taget_modules	q_proj,k_proj,v_proj,o_proj down_proj,gate_proj,up_proj

Table 9: LoRA configuration for LS-LLMs and IFT-LLMs

LoRA configuration item	Value
lora_r	32
lora_alpha	64
lora_dropout	0.05
lora_bias	none
lora_taget_modules	query, key, value, dense

Table 10: LoRA configuration for encoders

B Instruction format for fine-tuning decoder-only LLMs

Sample instruction for BC5CDR

Instruction:
please extract named entities and their type from the input sentence, all entity types are in options.
Options:
chemical,disease
Sentence:
The hypotensive effect of 100 mg / kg alpha-methyldopa was also partially reversed by naloxone .
Response:
alpha-methyldopa:chemical;naloxone:chemical;hypotensive:disease;

Sample instruction for CoNLL2003

Instruction:
please extract named entities and their type from the input sentence, all entity types are in options.
Options:
organization,miscellaneous,person,location
Sentence:
The European Commission said on Thursday it disagreed with German advice to consumers to shun British lamb until scientists determine whether mad cow disease can be transmitted to sheep .
Response:
European Commission:organization;German:miscellaneous;British:miscellaneous;

C Zero-shot NER using GPT-4.1 by OpenAI

The following is a sample code snippet of zero-shot NER for BC5CDR and CoNLL2003 using OpenAI's structured output.

```python
from openai import OpenAI
from pydantic import BaseModel
from typing import Literal

client = OpenAI()

class Entity_BC5CDR(BaseModel):
    disease: list[str]
    chemical: list[str]

class Entity_CoNLL2003(BaseModel):
    organization: list[str]
    miscellaneous: list[str]
    person: list[str]
    location: list[str]

def extract_entity(content: str, task: Literal["BC5CDR","CoNLL2003"]):

    response = client.responses.parse(
        model="gpt-4.1",
        input=[
                {
                    "role": "system",
                    "content": "You are an excellent linguist. Extract all the mentions that can be considered diseases or chemicals." if task == 'BC5CDR' else "You are an excellent linguist. Extract all the mentions that can be considered organizations, people, locations, or miscellaneous.",
                },
                {
                    "role": "user",
```

```
                    "content": f"{content}",
                },
            ],
        text_format=Entity_BC5CDR if task == 'BC5CDR' else Entity_CoNLL2003,
    )
    entities = response.output_parsed
    return entities

if __name__ == "__main__":
    entities = extract_entity(content, task_name)
```

D Error analysis

NER errors are classified into SPAN, TYPE, and DETECTION error categories, and each error category is further classified into finer error types. Errors are computed by comparing gold and predicted labels, each represented as a list of entity labels of the form (start, end, etype), where 'start' and 'end' are the beginning and end of the entity span, and 'etype' is the entity type.

In SPAN errors, if either 'start' or 'end' position matches, but the span length is not matched, they are classified as too_short or too_long. Whereas in the case of both 'start' and 'end' positions not matching, but the difference of span length is within two words, it is classified as boundary_off. The others are classified as completely_wrong. If span position matches, but 'etype' does not match, they are classified as correct_span_wrong_type.

BC5CDR	**DeBERTa**	**LS-FA-Llama3.1-8B**	**LS-CA-Llama3.1-8B**
Total Prediction	10502	10442	11383
Exact Matches	8705	9007	7247
Total Errors	1797	1435	4136
SPAN ERRORS			
boudary_off	5 (0.3%)	4 (0.3%)	1 (0.0%)
too_short	198 (11.0%)	195 (13.6%)	284 (6.9%)
too_long	244 (13.6%)	169 (11.8%)	138 (3.3%)
completely_wrong	7 (0.4%)	6 (0.4%)	10 (0.2%)
TYPE ERRORS			
correct_span_wrong_type	39 (2.2%)	14 (1.0%)	195 (4.7%)
DETECTION ERRORS			
false_positives	693 (38.6%)	633 (44.1%)	1574 (38.1%)
false_negatives	611 (34.0%)	414 (28.9%)	1934 (46.8%)

Table 11: Error analysis of DeBERTa, LS-FA-Llama3.1-8B, and LS-CA-Llama3.1-8B for BC5CDR

CoNLL2003	**DeBERTa**	**LS-FA-Llama3.1-8B**	**LS-CA-Llama3.1-8B**
Total Prediction	5800	5775	5912
Exact Matches	5177	5283	4219
Total Errors	623	492	1693
SPAN ERRORS			
boudary_off	0 (0.0%)	1 (0.2%)	0 (0.0%)
too_short	52 (8.3%)	47 (9.6%)	49 (2.9%)
too_long	75 (12.0%)	44 (8.9%)	29 (1.7%)
completely_wrong	0 (0.0%)	0 (0.0%)	0 (0.0%)
TYPE ERRORS			
correct_span_wrong_type	236 (37.9%)	176 (35.8%)	734 (43.4%)
DETECTION ERRORS			
false_positives	152 (24.4%)	127 (25.8%)	264 (15.6%)
false_negatives	108 (17.3%)	97 (19.7%)	617 (36.4%)

Table 12: Error analysis of DeBERTa, LS-FA-Llama3.1-8B, and LS-CA-Llama3.1-8B for CoNLL2003

A Multi-View Framework for Cross-Domain Nutrition Misinformation Detection in Social Media

Vishwaa Shah[1], Indika Kahanda[1], Andrea Arikawa[2], Asal Abbaszadeh[2], Richard Loftis[2]

[1]School of Computing, University of North Florida, Jacksonville, FL, USA
[2] Department of Nutrition & Dietetics, University of North Florida, Jacksonville, FL, USA

{N01458714, indika.kahanda}@unf.edu
{a.arikawa, n01481527, n01501201}@unf.edu

Abstract

Nutrition misinformation on social media often arises from selective interpretation of scientific evidence rather than outright falsehoods, making it difficult to detect. We introduce a curated, expert-annotated Instagram dataset focused on seed oils and omega-6, two domains characterized by contested dietary claims. We evaluate feature-based, embedding-based, and transformer-based models under in-domain and cross-domain settings. Results show strong in-domain performance across all models, with Sentence-BERT achieving the highest AUPRC (up to 0.96). However, performance drops substantially under cross-domain transfer, indicating limited robustness to topic shift. Analysis suggests that while contextual embeddings capture strong in-domain semantic signals, linguistically and psychologically grounded features are more stable under distribution shift. These findings highlight the value of combining semantic and interpretable linguistic signals for robust misinformation detection.

1 Introduction

Nutrition misinformation has emerged as a significant public health concern, driven in part by the rapid dissemination of unverified or misleading dietary claims on social media (Diekman et al., 2023). Prior research shows that online health information frequently contains inaccurate or contextually distorted statements that can shape public understanding and influence health-related decision-making (Tan et al., 2025). Although recent advances in large language models (LLMs) have improved automated detection of misleading content, reliable identification remains challenging due to the domain-specific and context-dependent nature of discourse (Belkhouribchia and Pen, 2025).

Social media platforms play a central role in shaping public nutrition narratives, but their scale and heterogeneity make systematic monitoring difficult (Garg and Sharma, 2022). Among these platforms, Instagram (https://www.instagram.com) is particularly influential due to its large user base and image-centric design driven by engagement (Dean, 2025). However, most existing misinformation datasets/studies focus on platforms such as Twitter and Facebook, leaving Instagram comparatively underexplored, particularly in nutrition-related contexts (Nguyen et al., 2022). This limits our understanding of how nutrition-related claims are communicated in environments where information is often condensed into short captions, hashtags, and visually supported narratives.

This gap is particularly evident in emerging debates in nutrition, such as on seed oils and omega-6. While dietary guidelines recognize vegetable oils as part of healthy dietary patterns when consumed appropriately (Dietary Guidelines for Americans, 2020), social media narratives frequently reframe them as inflammatory or harmful (Ahmad et al., 2025). Such content often does not rely on entirely fabricated claims, but instead on selective interpretation or reframing of established scientific evidence (Lee and Kurniawan, 2025), making detection more subtle and context-dependent.

From a computational perspective, previous work has shown that linguistic and psycholinguistic features can capture stylistic and rhetorical signals associated with misinformation (Garg and Sharma, 2022). In particular, psycholinguistic cues have been found useful for identifying emotional intensification and rhetorical framing in health-related misinformation (Mahbub et al., 2022). However, generalization across domains remains an open challenge, with performance often degrading when models are applied to unseen topics or distributions (Xarhoulacos et al., 2021).

In this work, we explore a hybrid framework for nutrition misinformation detection in Instagram posts. As outlined in Figure 1, we construct a curated expert-annotated dataset of Instagram captions focusing on seed oils and omega-6. We pro-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 326–341
July 3-4, 2026 ©2026 Association for Computational Linguistics

pose a multi-view representation that integrates linguistic complexity, rhetorical and stylistic features, psycholinguistic signals, and pretrained semantic embeddings. We evaluate both feature-based and transformer-based models under in-domain and cross-domain transfer settings to assess performance under domain shift.

Our experiments show that transformer-based models achieve strong in-domain performance, while feature-based models remain competitive and provide improved interpretability. However, all models exhibit performance degradation under cross-domain evaluation, particularly in ranking-based metrics such as AUPRC. These findings suggest that while both linguistic and neural representations capture useful signals, their ability to generalize across closely related nutrition topics needs further improvement. This highlights the value of combining interpretable linguistic signals with contextual embeddings for misinformation analysis.

2 Related Work

Health misinformation detection has been widely studied using natural language processing techniques, with transformer-based architectures and LLMs achieving strong performance in assessing the credibility of online content (Faruk, 2024; Tan et al., 2025). However, much of this work relies on general-purpose datasets or structured textual sources, which may not fully capture the informal, context-dependent nature of social media discourse, where health misinformation is often embedded in short, unstructured narratives (Belkhouribchia and Pen, 2025; Kauttonen et al., 2020).

A large portion of existing research has focused on platforms such as Twitter, Facebook, and YouTube (Yeung et al., 2022), while Instagram remains relatively underexplored despite its large user base and distinctive visual and caption-driven communication style (Nguyen et al., 2022). This is particularly important in health-related domains, where misinformation is often expressed through short captions, hashtags, and implicit claims embedded in multimodal content. Recent evidence suggests that nutrition misinformation is increasingly prevalent on Instagram, though systematic analysis remains limited due to the lack of curated datasets (Segado-Fernández et al., 2025).

Recurring nutrition misinformation narratives often center on dietary fats, particularly seed oils and omega-6 fatty acids (omega-6). These narratives frequently characterize such fats as harmful or pro-inflammatory, despite established dietary evidence supporting their role in healthy dietary patterns when consumed appropriately (Lee and Kurniawan, 2025; Ahmad et al., 2025). This gap between scientific consensus and public framing highlights the need for computational approaches that capture not only the content of claims but also their rhetorical and stylistic presentation.

Prior work in misinformation detection has demonstrated the usefulness of linguistic and stylistic features, including lexical diversity, readability, and syntactic complexity (Garg and Sharma, 2022). Misinformation is often associated with distinctive rhetorical patterns such as increased certainty, reduced hedging, and emotionally charged language (Mahbub et al., 2022). Hybrid approaches combining such features with dense embeddings, such as WELFake (Verma et al., 2021), have shown improvements in classification performance, though their effectiveness varies across domains and datasets (Khan et al., 2021).

Emotion and affective signals have also been widely studied as predictive indicators of misinformation. Prior studies suggest that misinformation content often exhibits stronger negative emotions such as fear, anger, and disgust compared to factual content (Farhoundinia et al., 2024). High-arousal emotional language has also been linked to reduced critical evaluation and increased susceptibility to misleading claims (Martel et al., 2020; Wurst et al., 2024), motivating their inclusion in computational models, although these signals may vary substantially across domains and contexts.

More recently, cross-domain misinformation detection has gained attention, particularly in health-related settings such as transferring models across disease topics (Goldani et al., 2025). These show that generalization across domains remains challenging, especially when moving between different thematic distributions or communication styles. However, most prior work focuses on disease-related misinformation, while broader lifestyle and nutrition domains remain relatively underexplored.

Despite these advances, existing approaches often treat linguistic features and neural representations separately, with limited systematic evaluation of their combination in domain-shift settings. This is particularly relevant in nutrition misinformation, where both interpretability and robustness are important. Motivated by this gap, we investigate a unified framework that integrates linguistic, rhetori-

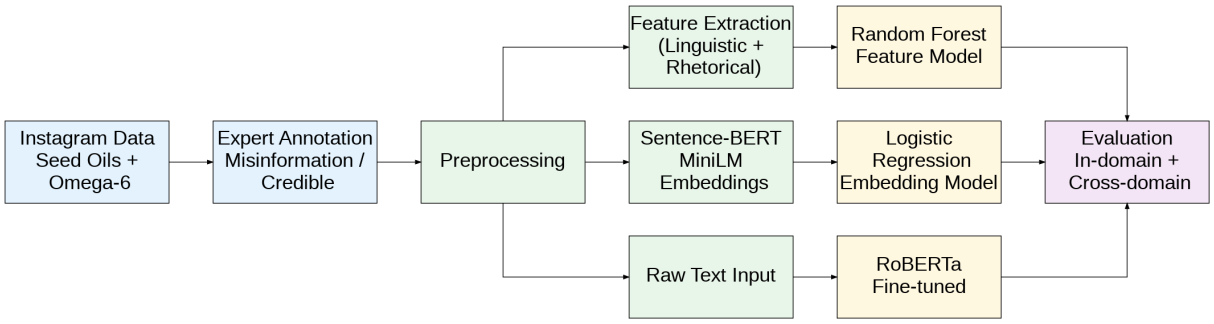


Figure 1: Overview of the proposed framework for nutrition misinformation detection. The pipeline includes data collection from Instagram, expert annotation, preprocessing, feature-based modeling (top branch), embedding-based modeling (middle branch), and transformer-based modeling (bottom branch). We evaluate robustness under domain shift, not just classification accuracy.

cal, and psycholinguistic features with transformer-based embeddings, and evaluate their effectiveness in both in-domain and cross-domain transfer.

3 Methodology

Figure 1 provides an outline of the proposed framework, summarizing the data collection, preprocessing, feature extraction, modeling approaches, and evaluation protocols used in this study.

3.1 Data Collection

A domain expert in nutrition identified relevant Instagram posts using the platform's native search interface. The sampling strategy was centered on two predefined fat-related topics: seed oils and omega-6. These topics were selected for focused analysis due to their frequent mischaracterization on social media as harmful or pro-inflammatory, despite established evidence supporting their role as beneficial unsaturated fats when consumed in appropriate dietary patterns, as reflected in the Dietary Guidelines for Americans 2020–2025 (Dietary Guidelines for Americans, 2020).

The keyword set was developed through expert consultation, iterative examination of nutrition-related content on social media, and preliminary testing of search outputs. Early exploration indicated that broad nutrition terms (e.g., "healthy eating") produced substantial amounts of irrelevant material, whereas targeted fat-specific queries yielded posts with clearer and more substantive health-related claims. The final keywords were chosen because they consistently surfaced high-engagement content and reflected areas where public misconceptions are especially prevalent.

Posts were considered eligible for inclusion if their captions explicitly referenced seed oils (e.g., soybean, sunflower, safflower, canola, or corn oil) or omega-6 in the context of nutrition, health, or disease-related claims. Only English-language posts were retained.

To mitigate the effects of algorithmic personalization, all searches were conducted using newly created Instagram accounts with no prior activity. Each account was used exclusively for one topic (seed oils or omega-6), ensuring that no cross-topic interaction influenced content retrieval. No engagement actions (e.g., likes, follows, or saves) were performed prior to or during data collection.

In addition, all search sessions were performed using a virtual private network (VPN) to minimize geographic and personalization biases. This approach was intended to approximate the content exposure of a new, unconditioned user encountering discussions related to these topics on Instagram.

Data collection took place between May 28, 2025 to June 4, 2025. Eligible posts were restricted to those published between January 1, 2020, and the end of the collection period. To focus on widely disseminated content, only posts from accounts with more than 5,000 followers were included.

All posts were publicly accessible at the time of collection, and no private accounts were accessed. Usernames and other identifying information were removed prior to analysis to preserve anonymity. Captions and associated metadata were extracted using the *Apify Instagram Post Scraper Actors*[1].

This study focuses exclusively on caption text rather than image or video content. This decision was made to enhance reproducibility and reduce ethical concerns, as textual data can be more readily anonymized, whereas visual content may contain identifiable individuals or sensitive information. By

[1] https://apify.com/actors, last accessed 06/04/2025.

isolating caption text, we attempt provide a lower-bound estimate of misinformation detectability, enabling clearer attribution of linguistic and rhetorical effects without visual confounds.

All preprocessing was implemented in Python within a Google Colab environment with fixed random seeds to ensure reproducibility. Preprocessing included removal of non-alphanumeric characters while preserving punctuation, normalization of whitespace, and retention of emojis as tokens. The annotated dataset and code are publicly available on Zenodo under a CC BY 4.0 license (https://doi.org/10.5281/zenodo.20247254).

3.2 Human Annotation

We define health misinformation in line with prior literature as "any health-related factual claim that contradicts established scientific consensus" (Sylvia Chou et al., 2020). In this work, scientific consensus is grounded in the 2020–2025 Dietary Guidelines for Americans (Dietary Guidelines for Americans, 2020). Content that contradicted or distorted these guidelines was annotated as *Misinformation*, whereas content that aligned with or did not conflict with the guidelines was labeled *Credible*. Consistent with prior public health misinformation research, our annotation framework evaluates claims relative to established evidence-based recommendations rather than isolated or emerging studies, and should therefore be interpreted as reflecting alignment with contemporaneous scientific consensus and public health guidance during the data collection period.

Three annotators with formal nutrition training independently labeled each post in two domains: seed oils and omega-6. The annotation team consisted of two graduate students in nutrition supervised by a PhD-trained nutrition faculty at University of North Florida. The annotation was guided by a detailed codebook derived from the Dietary Guidelines for Americans (see the Appendix A). To reduce potential bias, annotators worked independently, were blinded to each other's decisions, and reviewed posts in randomized order. The Inter-rater reliability (IRR) was calculated using Fleiss's κ ($= 0.814$), which accounts for agreement occurring by chance and indicates strong agreement (values above 0.80). Given this high agreement, we used majority voting without additional adjudication to preserve annotation independence. These consensus labels serve as the basis for the subsequent evaluation of model predictions.

Figure 2 presents the distribution of caption lengths by label across both domains. Overall, captions are relatively long, with mean lengths of 156.1 words (SD = 108.5) in the seed oil domain and 236.5 words (SD = 101.2) in the omega-6 domain, indicating that posts often require interpretation of extended context rather than isolated claims. In the seed oil domain, credible posts are longer on average (M = 179.5, SD = 113.9) than misinformation posts (M = 146.8, SD = 105.4), suggesting that misinformation content tends to be more concise in this domain. In contrast, in the omega-6 domain, caption lengths are highly similar across classes, with credible posts (M = 238.9, SD = 110.4) and misinformation posts (M = 235.8, SD = 98.9) showing nearly identical averages. Across both domains, caption lengths vary widely, with an overall mean of 193.6 words (SD = 112.4), reflecting substantial heterogeneity in how claims are presented.

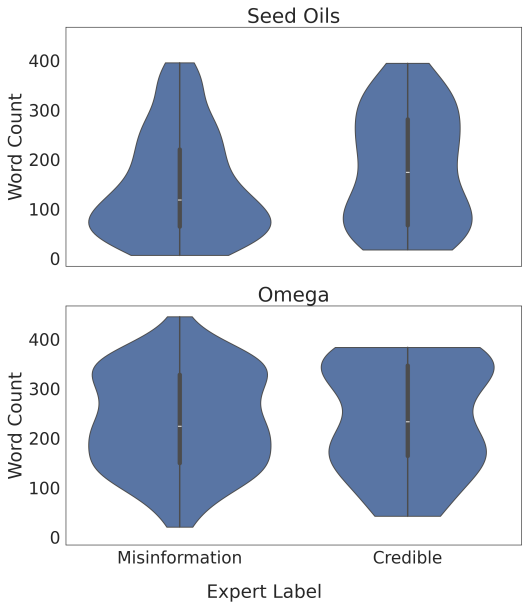


Figure 2: Caption length (word count) by expert label across seed oil and omega-6 domains. Captions in both domains are often long-form, requiring contextual interpretation beyond isolated claims.

The class distribution differs across domains but is consistently skewed toward misinformation. In the omega-6 dataset, 77.0% (114/148) of posts were labeled as *Misinformation*, while 23.0% (34/148) were labeled as *Credible*. In the seed oil dataset, 71.6% (121/169) of posts were labeled as *Misinformation*, compared to 28.4% (48/169) labeled as *Credible*. All captions were manually reviewed in full. Although some posts were brief or minimally informative, they were retained to

preserve the variability characteristic of real-world social media content.

3.3 Feature-Based Model

We develop a topic-agnostic feature-based framework to investigate how misinformation is linguistically and affectively distinguished from credible nutrition discourse. Instead of relying on surface lexical cues or topic-specific words, we focus on capturing how claims are framed across linguistic, rhetorical, and psychological dimensions.

3.3.1 Feature Space Construction

We construct a unified representation consisting of four complementary feature groups. Together, these four types of features form a comprehensive, topic-agnostic representation of linguistic style, persuasion strategy, and affective structure. A complete definition of all features is provided in Appendix B.

(1) Linguistic Complexity Features. We quantify structural and readability properties using `textstat` and `lexicalrichness`, including Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, lexical diversity, word count, and average word length. These features capture cognitive accessibility and structural complexity independent of topic.

(2) Rhetorical and Persuasive Features. We extract discourse-level signals using rule-based patterns to identify persuasive and stylistic markers. These include certainty and hedging expressions, rhetorical questions, exclamation usage, capitalization intensity, and citation or evidence indicators. These features model how claims are communicated rather than their content. Detailed rule-based patterns used for rhetorical feature extraction are provided in Appendix C.

(3) Affective Features. We use pretrained transformer models to extract sentiment and emotion signals from the transformers library (*cardiffnlp/twitter-roberta-base-sentiment-latest* and *j-hartmann/emotion-english-distilroberta-base*) to gauge the emotional tone of texts. These produce sentiment labels and emotion probability distributions. The full emotion label mapping and output structure of the classifier are provided in (Appendix D).

(4) Psychological Valence and Arousal Features. We operationalize affect using a lexicon-based approach grounded in the Research Domain Criteria (RDoC) framework (Insel et al., 2010), capturing Positive Valence (reward and approach motivation), Negative Valence (threat, loss, frustration), and Arousal/Regulatory systems (activation and intensity). These are combined with VADER (Hutto and Gilbert, 2014) sentiment scores to derive interpretable psychological dimensions reflecting emotional directionality, intensity, and motivational framing. Unlike categorical emotion labels, RDoC dimensions are designed to capture domain-independent affective processes such as threat sensitivity and approach motivation, making them well-suited for cross-topic misinformation analysis. The full lexicons used for RDoC feature extraction are provided in (Appendix E).

3.3.2 Feature-Based Classification Model

Random Forest achieved the best and most stable performance among several traditional machine learning classifiers, and was therefore selected as the final classifier. The model is implemented using `scikit-learn` and trained on the concatenated feature representation consisting of linguistic, rhetorical, and affective features.

Features are standardized prior to training, and model interpretability is examined using SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017). SHAP is used to quantify feature contributions at both global and instance levels, enabling analysis of which linguistic, rhetorical, and affective signals most strongly influence misinformation predictions.

SHAP values are computed using model-specific explainers: *TreeExplainer* for tree-based models, *LinearExplainer* for logistic regression, and *KernelExplainer* when required for non-linear models. We compute SHAP values on the full dataset using the final trained model and report mean absolute SHAP values for the positive class (misinformation) as global feature importance. We further compare SHAP-based rankings with built-in feature importance from tree-based models to assess consistency across interpretation methods.

3.4 Embedding-based model

We compute sentence-level embeddings using the Sentence-BERT model `all-MiniLM-L6-v2` (`https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`), which produces dense contextual representations of captions. These embeddings are used as features in a logistic regression classifier, allowing us to evaluate whether pretrained semantic representa-

tions improve generalization beyond handcrafted linguistic and psychological features. We select `all-MiniLM-L6-v2` due to its balance between computational efficiency and semantic performance, enabling fast inference while maintaining strong performance in sentence-level similarity tasks. This setup isolates the contribution of pretrained embeddings by using a simple linear classifier without additional modeling complexity.

3.5 Transformer-Based Model

We fine-tune `roberta-base` (Liu et al., 2019) for binary misinformation classification. RoBERTa is selected because it provides strong contextual language understanding and serves as a high-capacity baseline to evaluate whether deep semantic representations alone can capture misinformation patterns without explicit feature engineering. Captions are tokenized with a maximum length of 256 tokens. Approximately 39% of captions exceed the 256-token limit and are truncated during tokenization. Truncation primarily affects longer narrative-style posts, whereas shorter captions remain fully preserved. The model is trained using a learning rate of 2×10^{-5}, batch size of 16, weight decay of 0.01, and early stopping based on validation set performance.

3.6 Evaluation: In-Domain & Cross-Domain

We evaluate models under two settings: **In-Domain Evaluation:** For each dataset (omega-6 and seed oils), we perform stratified 5-fold cross-validation, ensuring balanced class distributions across folds. This procedure is applied consistently across all three modeling approaches. **Cross-Domain Evaluation:** We conduct bidirectional transfer experiments to evaluate cross-domain generalization. Specifically, models are trained on the full source-domain dataset and evaluated on the target-domain dataset without any adaptation. We consider omega-6$\rightarrow$ seed (training on omega-6, testing on seed oils) and seed $\rightarrow$ omega-6 (training on seed oils, testing on omega-6). This setup evaluates robustness under realistic domain shift, where topic-specific cues may not transfer across domains.

All experiments are conducted with a fixed random seed of 42 to ensure reproducibility. In the cross-domain setting, no information from the target domain is used during training or model selection, thereby preventing any form of data leakage.

3.7 Evaluation Metrics

We use evaluation metrics suited for imbalanced classification and misinformation detection, including precision, recall, and F1-score (for the misinformation class), as well as the area under the precision–recall curve (AUPRC). F1-score is the harmonic mean of precision and recall, providing a balanced measure under class imbalance. AUPRC is a threshold-independent metric that summarizes ranking performance and is particularly informative when the positive class is of primary interest. Although the class distribution is moderately imbalanced, we use AUPRC because it evaluates ranking performance independent of decision thresholds and provides a more informative measure of classifier behavior under non-uniform class distributions.

4 Results

4.1 In-Domain Performance

All models perform well on the misinformation class, with consistent trends across both domains (Table 1). To account for class imbalance, we compare against baselines based on majority-class and prevalence, which produce identical performance with AUPRC values of 0.77 (seed oils) and 0.70 (omega-6), reflecting the distributions of the underlying labels. All proposed models consistently outperform these baselines under in-domain settings and achieve higher ranking performance than baseline under most cross-domain transfers, indicating that improvements are not driven solely by class imbalance.

Sentence-BERT (MiniLM) achieves the highest overall performance, reaching 0.96 AUPRC on seed oils and 0.92 on omega-6, outperforming both RoBERTa and the feature-based Random Forest model in metrics. RoBERTa achieves perfect recall (1.00) across both datasets, but with lower precision (0.77 on omega-6, 0.72 on seed oils), resulting in reduced F1. This indicates a consistent tendency to over-predict the positive class.

The Random Forest model remains competitive, reaching F1 scores of 0.87 (omega-6) and 0.86 (seed oils). Despite using only engineered linguistic and rhetorical features, it performs close to neural approaches, indicating that structured features retain strong predictive value.

4.2 Cross-Domain Generalization

Across all models, AUPRC drops under cross-domain transfer (Table 2), indicating reduced rank-

Model	Domain	Prec	Rec	F1	AUPRC
MiniLM	omega-6	0.90	0.89	0.90	0.92
MiniLM	seed oils	0.87	0.89	0.88	0.96
RoBERTa	omega-6	0.77	1.00	0.87	0.86
RoBERTa	seed oils	0.72	1.00	0.83	0.84
Random Forest	omega-6	0.80	0.96	0.87	0.90
Random Forest	seed oils	0.79	0.95	0.86	0.84

Table 1: In-domain performance.

Train → Test	Model	Prec	Rec	F1	AUPRC
omega-6→ seed	MiniLM	0.72	1.00	0.83	0.74
omega-6→ seed	RoBERTa	0.72	1.00	0.83	0.69
omega-6→ seed	RF	0.72	1.00	0.83	0.74
seed → omega-6	MiniLM	0.77	0.95	0.85	0.86
seed → omega-6	RoBERTa	0.77	1.00	0.87	0.82
seed → omega-6	RF	0.77	0.90	0.83	0.80

Table 2: Cross-domain performance.

ing performance when models are applied to unseen topics. For example, MiniLM decreases from 0.92 in-domain on omega-6 to 0.74 when transferred from omega-6 to seed oils, while RoBERTa decreases from 0.86 to 0.69 in the same setting. The degradation is more pronounced for the omega-6→seed transfer direction (Figure 3).

Sentence-BERT (all-MiniLM-L6-v2) achieves the strongest overall in-domain performance and competitive cross-domain AUPRC across both transfer settings, suggesting relatively strong generalization ability compared to the other approaches.

The Random Forest model also demonstrates competitive cross-domain robustness, particularly in the seed→omega-6 setting, where it shows only a modest decrease in AUPRC (0.84 to 0.80). This suggests that topic-agnostic linguistic and rhetorical features may capture more transferable misinformation signals. In particular, rhetorical features (e.g., certainty markers) may generalize more consistently across topics than dense semantic representations.

4.3 Linguistic and Rhetorical Differences

Readability analysis indicates that captions across both classes are moderately complex, with no strong separation in Flesch Reading Ease scores. In the omega-6 domain, credible posts exhibit slightly higher readability (M = 52.95, SD = 11.76) than misinformation (M = 49.26, SD = 15.19). A similar pattern is observed in the seed oil domain (credible: M = 56.57, SD = 15.98; misinformation: M = 51.40, SD = 23.15). However, the substantial overlap and variance suggest that readability is not a primary distinguishing factor.

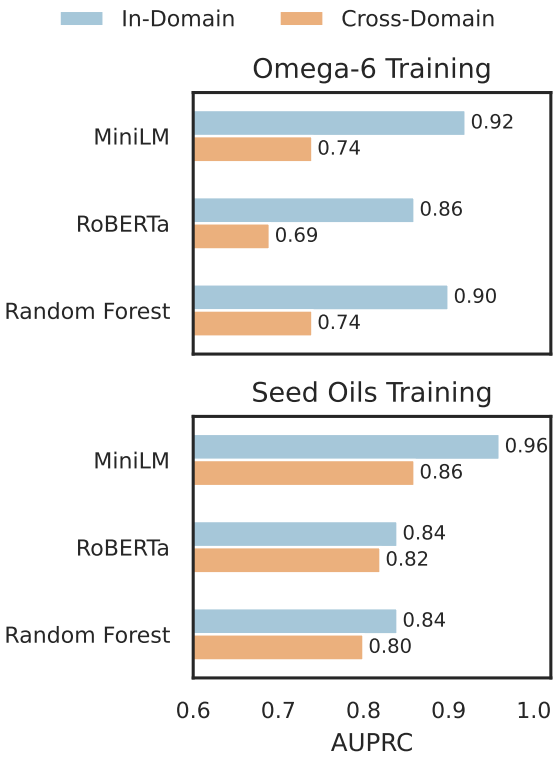


Figure 3: AUPRC comparison across in-domain and cross-domain transfer for omega-6 and seed oil datasets.

Clearer distinctions emerge in rhetorical structure (Figure 4). Across both domains, misinformation consistently exhibits lower levels of hedging, indicating a more assertive and less qualified communication style, while credible posts show more frequent use of hedging and exploratory language.

Misinformation is also characterized by stronger reliance on persuasive and heuristic framing, including more frequent appeal-to-nature arguments and the presence of conspiracy-related cues, which are largely absent from credible posts. In contrast, credible posts show slightly higher levels of exploratory questioning, suggesting a more open-ended or deliberative tone. Misinformation additionally exhibits more emphatic signaling (e.g., exclamation usage and capitalization), particularly in the seed oil domain. Overall, the differences are primarily rhetorical rather than lexical or readability-based, with misinformation distinguished by higher certainty, stronger framing devices, and reduced epistemic caution.

4.4 Affective/Psychological Feature Analysis

Affective analysis reveals a nuanced pattern suggesting affective simplification in misinformation. In both domains, misinformation shows a more con-

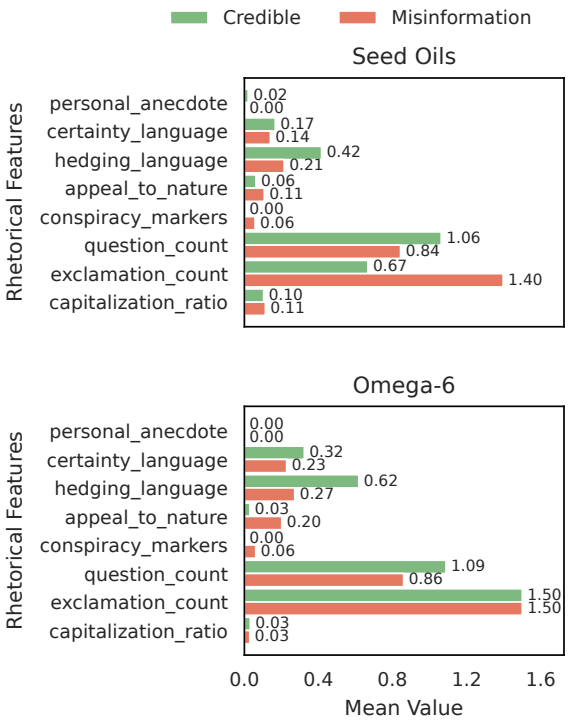


Figure 4: Rhetorical and linguistic feature distributions across credible and misinformation classes.

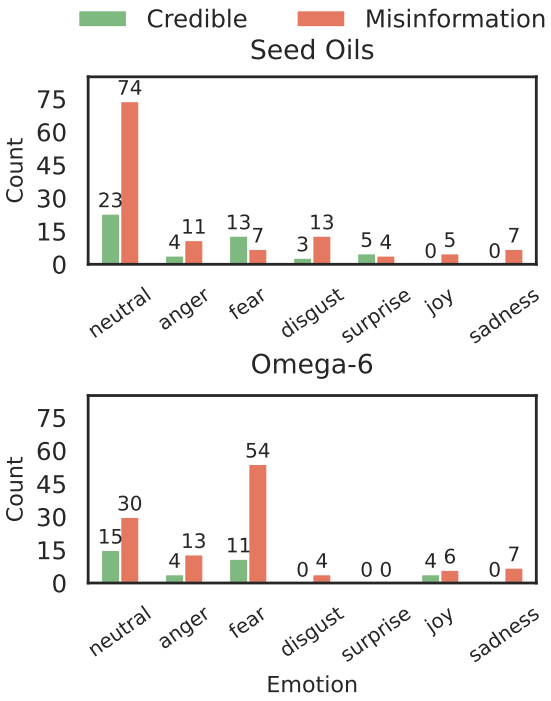


Figure 5: Emotion distribution across credible and misinformation posts in the seed oil and omega-6 domains.

centrated emotional profile, while credible content exhibits a less skewed distribution of affect (Figure 5). In the seed oil domain, misinformation is dominated by neutral affect with secondary anger and fear, whereas credible posts span a broader set of emotions, including fear and disgust. In the omega-6 domain, misinformation is strongly concentrated in fear-related language, while credible content distributes more evenly across negative and mixed emotions. Overall, misinformation is characterized by reduced emotional variability, while credible content shows a more heterogeneous affective distribution.

Across both domains, misinformation captions show lower scores across psychological dimensions, including valence, emotional intensity, and reward/threat systems, indicating reduced emotional richness. These findings suggest that misinformation is characterized not by emotional amplification, but by simplified and less varied affective expression, which may increase perceived clarity and directness.

4.5 Feature Importance and Interpretability

A total of 55 features were extracted per domain. SHAP analysis was conducted over the entire feature space without feature selection to provide an unbiased estimate of the importance of the feature. Figure 6 presents the top 10 features ranked by mean absolute SHAP values.

Across both domains, the most influential features are primarily psychological and affective. Top features include RDoC-based measures of threat and valence, emotion-neutral signals, and sentiment-related features. Rhetorical features, including appeal to nature and question frequency, also contribute to model predictions, highlighting the importance of communicative framing.

The consistency of these features across domains indicates that misinformation is characterized by stable, topic-agnostic signals, including reduced hedging, lower emotional complexity, and increased reliance on persuasive heuristics. These patterns help explain model behavior under domain shift: models likely maintain high recall by capturing generalizable stylistic signals, but may experience reduced AUPRC due to the loss of topic-specific semantic information.

5 Conclusion

In this work, we presented a multi-view framework for detecting nutrition misinformation in Instagram posts, focusing on seed oils and omega-6. We introduced a curated, expert-annotated dataset and eval-

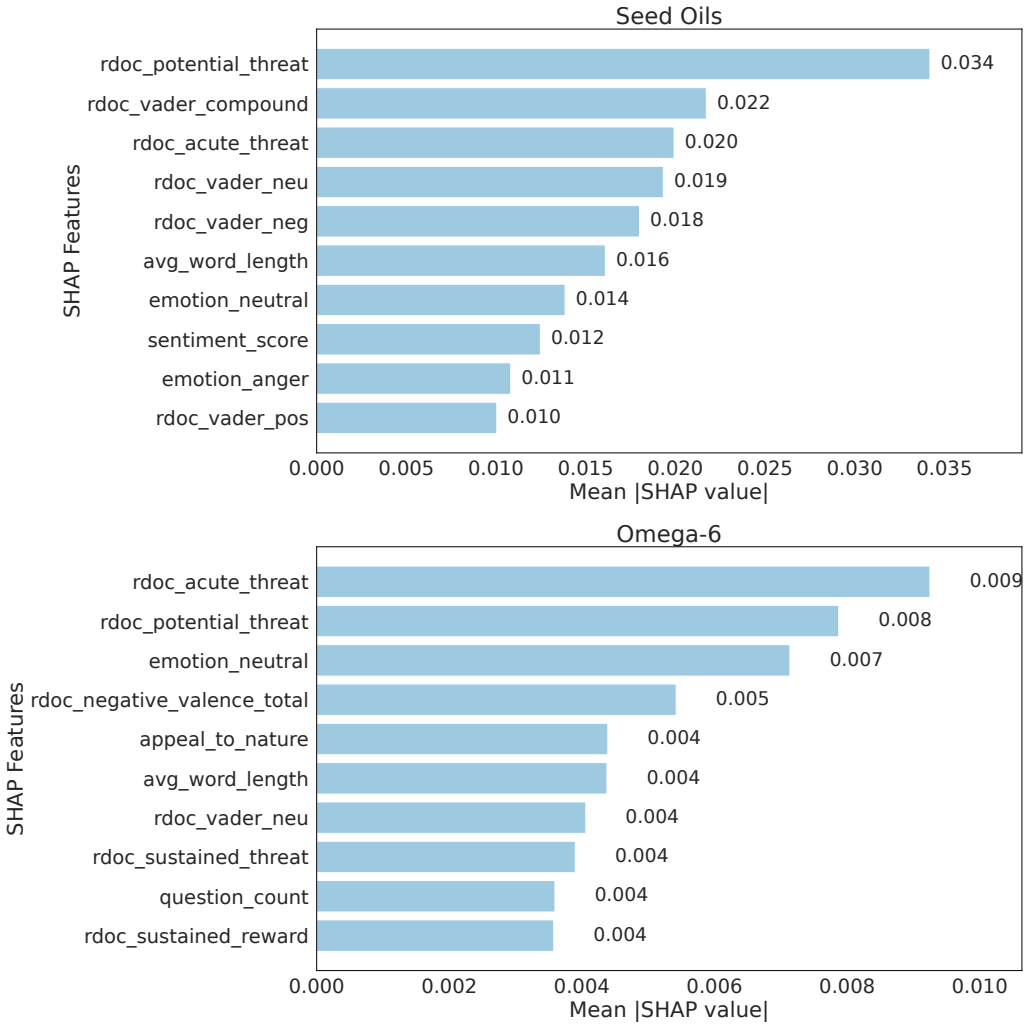


Figure 6: Top 10 features ranked by mean absolute SHAP values for the Random Forest model.

uated both feature-based and transformer-based models under in-domain and cross-domain transfer. Our results show that Sentence-BERT achieves the strongest in-domain performance, while RoBERTa prioritizes recall at the expense of precision. Notably, a feature-based Random Forest model leveraging linguistic, rhetorical, and psycholinguistic features performs competitively, highlighting the value of interpretable representations alongside deep semantic models.

We observe substantial performance degradation under cross-domain evaluation, even between closely related nutrition topics, underscoring limited generalization under domain shift. Beyond predictive performance, our analysis reveals consistent stylistic and affective patterns in misinformation, characterized by rhetorical certainty, reduced hedging, and simplified emotional expression. Feature importance analysis further confirms that psychological and affective signals are among the most informative predictors. Overall, these findings highlight the complementary strengths of interpretable linguistic features and contextual embeddings for robust and explainable nutrition misinformation detection, and motivate future work on multimodal and cross-platform generalization.

Limitations

The dataset used in this study is relatively small and restricted to two closely related nutrition topics, namely seed oils and omega-6. While this focused design allows for controlled analysis of domain shift, it may limit the broader applicability of the findings to other areas of nutrition or health misinformation.

The analysis is also limited to English-language Instagram captions, which excludes multilingual content and may not capture cultural or linguistic variations in how nutrition misinformation is expressed globally. In addition, only textual caption

data is considered, while Instagram is inherently multimodal and includes images, videos, and user engagement signals that may provide important contextual cues for misinformation detection.

Although the annotation process was guided by detailed expert-defined guidelines and inter-annotator agreement was measured, labeling misinformation inevitably involves a degree of subjectivity. Despite efforts to reduce bias through training and independent annotation, some ambiguity may persist, particularly in borderline cases where claims are partially supported or selectively framed. We also acknowledge that nutrition science is an evolving field and that dietary guidance may change over time as new evidence emerges. As with all evidence-based guidelines, the DGA reflects the prevailing scientific consensus at a particular point in time and is therefore subject to revision as the evidence base develops. Accordingly, the annotation framework used in this study should be interpreted as reflecting alignment with contemporaneous public health guidance during the data collection period rather than immutable scientific ground truth.

Finally, while the proposed feature set captures linguistic, rhetorical, and psycholinguistic signals, it is still based on predefined assumptions about informative cues. Transformer-based models are also constrained by the relatively small dataset size and may benefit from larger-scale or domain-adaptive training. Cross-domain evaluation is limited to topic transfer within Instagram, and broader cross-platform studies would be necessary to fully assess real-world generalization.

Ethical Considerations

This study uses only publicly available Instagram content and analyzes caption text without collecting private or sensitive user data. All identifiers were removed to ensure anonymity, and the dataset is fully de-identified. The models developed in this work are intended for research purposes and should be interpreted as probabilistic indicators rather than definitive judgments of truth or falsehood. Our annotation framework operationalizes misinformation relative to contemporaneous evidence-based public health guidance, specifically the 2020–2025 Dietary Guidelines for Americans and the accompanying Scientific Report of the Dietary Guidelines Advisory Committee. These guidelines are developed through systematic evidence review and expert synthesis of the nutrition literature and are intended to reflect prevailing scientific consensus at the time of publication. Consistent with prior public health misinformation research, we evaluate claims relative to established evidence-based recommendations rather than isolated or emerging studies. We acknowledge that nutrition science continues to evolve and that there may be some scientific disagreement; however, individual conflicting findings do not necessarily outweigh the broader consensus reflected in guideline-based recommendations. Care is required when applying misinformation detection systems in real-world settings, as misclassification risks may disproportionately affect nuanced or evolving scientific discussions. This work aims to support, rather than replace, expert evaluation of nutrition-related content.

Acknowledgments

We thank the nutrition experts, Dr. Charlotte Martin, DCN, RDN, and Dr. Alan Flanagan, PhD, for their valuable contributions to screening/extracting/annotating the original Instagram posts. The authors also gratefully acknowledge the School of Computing, College of Computing, Engineering and Construction, and Graduate School at the University of North Florida for their support and funding, which made this research possible.

References

Sharique Ahmad, Subuhi Anwar, and Pushpendra D. Pratap. 2025. The cod liver oil insights: From folklore to facts. *International Journal of Research - GRANTHAALAYAH*, 13(8):95–109.

Jamal Belkhouribchia and Joeri Jan Pen. 2025. Large language models in clinical nutrition: an overview of its applications, capabilities, limitations, and potential future prospects. *Frontiers in Nutrition*, 12:1635682.

Brian Dean. 2025. Instagram demographic statistics: How many people use instagram in 2024? Accessed: 2025-04-25.

Connie Diekman, Camille D Ryan, and Tracy L Oliver. 2023. Misinformation and disinformation in food science and nutrition: impact on practice. *The Journal of Nutrition*, 153(1):3–9.

Dietary Guidelines for Americans. 2020. Dietary guidelines for americans, 2020–2025 and online materials. https://www.dietaryguidelines.gov/resources/2020-2025-dietary-guidelines-online-materials.

Reyhaneh Farhoundinia, Daniel Munro, and Mark de Rond. 2024. Emotions unveiled: detecting COVID-19 fake news on social media. *Humanities and Social Sciences Communications*, 11(1):1–14.

Tanjim Bin Faruk. 2024. Evaluating the performance of large language models in scientific claim detection and classification. https://arxiv.org/abs/2412.16486.

Shubham Garg and Sonal Sharma. 2022. Linguistic features based framework for automatic fake news detection. *Computers & Industrial Engineering*, 172:108432.

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2025. Fighting misinformation in health news: Dcnn-capsnet for cross-domain health misinformation detection. *ACM Transactions on Intelligent Systems and Technology*.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Thomas R. Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–756.

Janne Kauttonen, Jenni Hannukainen, Pia Tikka, and Jyrki Suomala. 2020. Predictive modeling for trustworthiness and other subjective text properties in online nutrition and health communication. *PloS one*, 15(8):e0237144.

Jahanzaib Zain Khan, Basheer Alam, and Youngmoon Lee. 2021. An approach utilizing linguistic features for fake news detection. In *Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1–6. IEEE.

Kevin Lee and Keszya Kurniawan. 2025. Are seed oils the culprit in cardiometabolic and chronic diseases? a narrative review. *Nutrition Reviews*, 83(7):e2106–e2112.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Sinthia Mahbub, Eric Pardede, and A. S. M. Kayes. 2022. COVID-19 rumor detection using psycholinguistic features. *IEEE Access*, 10:117530–117543.

Cameron Martel, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):1–20.

Van Nguyen, Luke Testa, Andrea L Smith, Louise A Ellis, Adam G Dunn, Jeffrey Braithwaite, and Mitchell Sarkies. 2022. Unravelling the truth: Examining the evidence for health-related claims made by naturopathic influencers on social media–a retrospective analysis. *Health Promotion Perspectives*, 12(4):372.

Sergio Segado-Fernández, Beatriz Jiménez-Gómez, Pedro Jesús Jiménez-Hidalg, María del Carmen Lozano-Estevan, and Iván Herrera-Peco. 2025. Disinformation about diet and nutrition on social networks: a review of the literature. *Nutrición Hospitalaria*, 42(2).

Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. 2020. Where we go from here: health misinformation on social media.

Dongmei Tan, Yi Huang, Ming Liu, Ziyu Li, Xiaoqian Wu, and Cheng Huang. 2025. Identification of online health information using large pretrained language models: Mixed methods study. *Journal of Medical Internet Research*, 27:e70733.

Pawan Kumar Verma, Priyanka Agrawal, Isabelle Amorim, and Radu Prodan. 2021. WELFake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Christiane Wurst, Paul Sauseng, Shruthi Prasad, Anchali Trakulp, Cameron Martel, and David Rand. 2024. Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance. *Cognitive Research: Principles and Implications*, 9(1):1–18.

Constantinos-Giovanni Xarhoulacos, Argiro Anagnostopoulou, George Stergiopoulos, and Dimitris Gritzalis. 2021. Misinformation vs. situational awareness: The art of deception and the need for cross-domain detection. *Sensors*, 21(16):5496.

Andy Wai Kan Yeung, Anela Tosevska, Elisabeth Klager, Fabian Eibensteiner, Christos Tsagkaris, Emil D Parvanov, Faisal A Nawaz, Sabine Völkl-Kernstock, Eva Schaden, Maria Kletecka-Pulker, and 1 others. 2022. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. *Journal of Medical Internet Research*, 24(1):e28152.

Appendix

A Annotation Guidelines

A.1 Purpose of These Guidelines

These guidelines ensure that all human coders classify social media nutrition posts consistently using:

- A two-level accuracy scale (0 or 1)

- Rules aligned with the *Dietary Guidelines for Americans, 2020–2025 (DGA)* and the *2020 DGAC Scientific Report*

Coders should reference only the content presented in the post and apply the rules exactly as described.

A.2 What Coders Evaluate

For each post, coders assign:

1. Accuracy classification (1 or 0)
2. An optional brief justification (2–3 sentences)

Coders do **not** evaluate:

- Creator's biography
- Hashtags
- Engagement metrics
- Intent or motivation
- User comments
- Information not shown in the post itself

A.3 Evidence Standard

Coders must evaluate claims solely against:

- *Dietary Guidelines for Americans 2020–2025*
- *DGAC Scientific Report (2020)*

Coders should **not** use:

- Personal knowledge
- Other nutrition guidelines
- Mechanistic physiology not discussed in the DGA/DGAC
- External sources, papers, or reputable organizations
- Internet searches

A.4 The Two-Level Accuracy Classification

1 - Accurate or Mostly Accurate A post is **Accurate** when:

- All nutrition claims fully align with DGA/DGAC
- Any minor inaccuracies do not change the overall message

Examples of Accurate Content

- Saying omega-6 fats are not inflammatory
- Correcting misinformation about seed oils

A post is **Mostly Accurate** when:

- It is evidence-aligned overall
- It contains some non-trivial inaccuracies, but the main message remains correct

Typical patterns

- Slight exaggeration without reversing the evidence

0 - Mostly Inaccurate or Inaccurate A post is **Mostly Inaccurate** when:

- It contains some correct information
- Misleading or incorrect claims dominate

Typical patterns

- Mixing evidence-based advice with unsupported mechanistic claims
- Overstating inflammation, hormone effects, oxidation, or fear-based messages about oils
- Encouraging elimination of seed oils while acknowledging they contain essential fats

A post is **Inaccurate** when:

- The core message contradicts DGA/DGAC
- Accurate statements, if present, are trivial

Examples

- "Butter is heart healthy; seed oils cause disease."
- "Seed oils are toxic poisons that cause inflammation."

A.5 Reasoning Requirements (2–3 Sentences)

Coders may provide a short justification for doubtful annotations. The justification must:

- Evaluate the overall message rather than line-by-line details
- Reference only areas covered by the DGA/DGAC
- Not mention excluded content (e.g., "the DGA doesn't cover . . . ")
- Avoid SEO phrases or quoting the creator directly
- Clearly explain why the classification was chosen

A.6 Special Situations

Mixed-accuracy posts Code based on the **overall message**, not on isolated statements.

Mechanistic claims (inflammation, oxidation, hormones) If the mechanism is not discussed in the DGA:

- Ignore the mechanism itself
- Evaluate whether the dietary conclusion contradicts DGA guidance

Reactive content If a post stitches another creator's video:

- Ignore the introductory clip
- Code only the creator's commentary

B Complete Feature Definitions

This section provides formal definitions of all features used in the feature-based model. Features are grouped into linguistic, rhetorical, affective (emotion and sentiment), psychological (RDoC-based), and interaction features. All continuous features are standardized prior to model training.

B.1 Linguistic Complexity Features

word_count Total number of words in the caption.

avg_word_length Average number of characters per word.

lexical_diversity Ratio of unique words to total words (type-token ratio).

flesch_reading_ease Readability score estimating ease of comprehension (higher values indicate easier text).

flesch_kincaid_grade Estimated U.S. grade level required to understand the text.

gunning_fog Readability index estimating years of formal education required to understand the text.

B.2 Rhetorical and Persuasive Features

certainty_language Count or binary indicator of certainty expressions (e.g., "always", "never", "prove", "guarantee").

hedging_language Count or binary indicator of uncertainty expressions (e.g., "may", "might", "possibly", "perhaps").

question_count Number of question marks in the text.

exclamation_count Number of exclamation marks in the text.

capitalization_ratio Proportion of uppercase characters relative to total characters.

appeal_to_nature Indicator for naturalistic fallacy framing (e.g., "natural is better", "chemical-free implies safe").

conspiracy_markers Indicator/count of conspiracy-related phrases (e.g., "big pharma", "cover-up", "they don't want you to know").

has_personal_anecdote Binary feature indicating presence of first-person experiential claims (e.g., "I tried", "my experience", "I noticed").

B.3 Interaction Features

hedge_x_exclaim Interaction between hedging intensity and exclamation usage.

certainty_x_caps Interaction between certainty language and capitalization ratio.

anecdote_x_exclaim Interaction between personal anecdote presence and exclamation usage.

word_count_x_lexical_div Interaction between text length and lexical diversity.

B.4 Emotion Features (Transformer-Based Model)

Emotion features are probability outputs from a pretrained DistilRoBERTa emotion classifier.

emotion_anger Probability of anger-related emotion.

emotion_fear Probability of fear or threat-related emotion.

emotion_disgust Probability of disgust or aversion.

emotion_joy Probability of positive affect or happiness.

emotion_sadness Probability of sadness or loss-related affect.

emotion_surprise Probability of surprise or unexpectedness.

emotion_neutral Probability of neutral or informational tone.

B.5 Sentiment Features (VADER-Based)

sentiment_score Overall sentiment polarity score ranging from negative to positive.

sentiment_positive Proportion of tokens contributing to positive sentiment.

sentiment_negative Proportion of tokens contributing to negative sentiment.

sentiment_neutral Proportion of tokens contributing to neutral sentiment.

rdoc_vader_pos Positive sentiment score mapped into RDoC positive valence space.

rdoc_vader_neg Negative sentiment score mapped into RDoC negative valence space.

rdoc_vader_neu Neutral sentiment score from VADER.

rdoc_vader_compound Normalized aggregate sentiment score combining all polarity signals.

B.6 RDoC-Based Psychological Features

B.6.1 Valence Features

rdoc_positive_valence_total Aggregate positive appraisal, reward, and pleasure-related language.

rdoc_negative_valence_total Aggregate threat, loss, and negative appraisal language.

rdoc_valence_dimension Net valence score computed as positive minus negative valence.

B.6.2 Arousal Features

rdoc_high_arousal High-energy emotional activation (urgency, excitement, intensity).

rdoc_low_arousal Low-energy emotional states (calmness, fatigue, neutrality).

rdoc_arousal_dimension Overall emotional activation level combining arousal signals.

rdoc_arousal_total Total magnitude of arousal-related activation.

rdoc_arousal_net Net difference between high and low arousal activation.

B.6.3 Threat and Reward System Features

rdoc_acute_threat Immediate or short-term threat-related language (fear, danger, urgency).

rdoc_potential_threat Uncertain or probabilistic threat framing.

rdoc_sustained_threat Chronic or long-term threat framing.

rdoc_loss Language indicating deprivation, harm, or negative outcomes.

rdoc_frustration Expressions of anger, irritation, or blocked goals.

rdoc_reward_expectancy Anticipation of positive outcomes or benefits.

rdoc_sustained_reward Stable or ongoing reward-related language.

rdoc_approach_motivation Goal-directed or desire-driven action language.

rdoc_reward_responsiveness Sensitivity or responsiveness to reward signals.

B.6.4 Composite Psychological Features

rdoc_emotional_intensity Overall magnitude of emotional expression derived from valence and arousal strength.

rdoc_emotional_complexity Entropy-based measure of diversity in emotional signals.

rdoc_threat_reward_ratio Ratio of threat-related to reward-related language intensity.

rdoc_valence_arousal_interaction Interaction between valence and arousal dimensions.

C Rhetorical and Persuasive Pattern Rules

We use rule-based pattern matching to extract rhetorical and persuasive features from text.

Certainty expressions:

- always, never, guarantee, prove, destroy

Hedging expressions:

- may, might, possibly, perhaps

Additional rhetorical markers:

- Question count: number of "?" in text
- Exclamation count: number of "!" in text
- Capitalization ratio: proportion of uppercase characters
- Citation indicators: presence of study, research, http, www, or year (e.g., 2020)

Conspiracy-related markers (used in exploratory analysis):

- terms such as "big pharma", "they don't want", "hidden", "cover-up"

D Emotion Model Output Mapping

Emotion features are extracted using a pretrained DistilRoBERTa emotion classifier. The model outputs probability distributions over the following emotion categories:

- anger
- disgust
- fear
- joy
- sadness
- surprise
- neutral

For each text, we store the probability score for each emotion class. These are used as continuous affective features in downstream classification.

E RDoC Lexicons for Valence and Arousal Features

We use manually constructed lexicons grounded in the Research Domain Criteria (RDoC) framework to operationalize psychological valence and arousal dimensions. These lexicons are used for feature extraction in the main analysis.

Positive Valence (Reward / Approach Motivation)

- reward_words: amazing, awesome, excellent, fantastic, great, wonderful, love, enjoy, happy, joy, excited, thrilled, delighted, perfect, best, incredible, outstanding, brilliant, superb, beautiful, gorgeous, lovely, pleasant, fabulous
- reward_expectancy: hope, expect, anticipate, look forward, optimistic, confident, believe, promise, guarantee, ensure, assure, certain
- approach_motivation: want, need, desire, wish, crave, seek, pursue, strive, goal, achieve, succeed, win, gain, obtain, acquire, try, attempt, effort, motivate, drive, determined

- reward_sustained: satisfaction, fulfillment, gratification, contentment, pleasure, reward, benefit, advantage, profit, success, achievement

Negative Valence (Threat / Loss / Frustration)

- acute_threat: fear, afraid, scared, terrified, frightened, panic, terror, alarmed, worried, anxious, nervous, dread, horror, dangerous, threat, risk, harm, danger, peril, hazard
- potential_threat: uncertain, unsure, doubt, suspicious, wary, cautious, concerned, apprehensive, uneasy, uncomfortable, tense, may, might, possibly, perhaps, suggest, potentially, risk
- sustained_threat: stress, chronic, ongoing, persistent, constant, continual, always, never-ending, relentless, unending, perpetual
- loss: lose, lost, loss, missing, gone, disappear, vanish, deprive, lack, without, absent, void, empty, sad, grief, mourn, regret, sorrow, heartbreak
- frustration: frustrate, annoyed, irritated, angry, mad, furious, rage, upset, disturbed, agitated, provoked, outraged, failed, failure, unsuccessful, denied, rejected, blocked

Arousal / Regulatory Systems

- high_arousal: intense, extreme, powerful, strong, overwhelming, explosive, urgent, critical, emergency, immediate, sudden, shock, surprise, astonish, startle, energy, active, lively, dynamic, vigorous, vibrant
- low_arousal: calm, peaceful, relaxed, tranquil, serene, quiet, slow, gentle, soft, mild, subtle, tired, fatigue, weary, exhausted, drained, lethargic, boring, dull, plain, mundane
- intensity_modifiers: very, extremely, incredibly, absolutely, totally, completely, utterly, highly, deeply, truly, so, quite, really

F Full SHAP Feature Importance Rankings

Seed Oils Domain
Feature
rdoc_acute_threat
rdoc_negative_valence_total
rdoc_emotional_intensity
rdoc_valence_arousal_interaction
emotion_sadness
emotion_disgust
rdoc_valence_dimension
rdoc_positive_valence_total
rdoc_threat_reward_ratio
emotion_fear
capitalization_ratio
rdoc_arousal_dimension
gunning_fog
rdoc_arousal_total
exclamation_count
rdoc_approach_motivation
emotion_surprise
rdoc_arousal_net
word_count_x_lexical_div
flesch_kincaid_grade
sentiment_score
hedging_language
rdoc_emotional_complexity
word_count
flesch_reading_ease
rdoc_potential_threat
emotion_neutral
rdoc_vader_compound
emotion_joy
rdoc_intensity_modifiers
lexical_diversity
avg_word_length
rdoc_vader_pos
rdoc_high_arousal
question_count
rdoc_vader_neg
emotion_anger
rdoc_vader_neu
rdoc_sustained_threat
rdoc_reward_expectancy
rdoc_loss
rdoc_sustained_reward
rdoc_reward_responsiveness
sentiment_neutral
hedge_x_exclaim
certainty_x_caps
conspiracy_markers
appeal_to_nature
has_personal_anecdote
sentiment_negative
certainty_language
rdoc_low_arousal
sentiment_positive
anecdote_x_exclaim
rdoc_frustration

Omega-6 Domain
Feature
rdoc_acute_threat
rdoc_potential_threat
emotion_neutral
rdoc_negative_valence_total
avg_word_length
appeal_to_nature
emotion_anger
rdoc_vader_neu
rdoc_sustained_threat
question_count
gunning_fog
rdoc_vader_neg
rdoc_sustained_reward
certainty_language
flesch_kincaid_grade
rdoc_high_arousal
hedging_language
rdoc_approach_motivation
rdoc_frustration
capitalization_ratio
rdoc_reward_responsiveness
rdoc_vader_pos
rdoc_valence_arousal_interaction
sentiment_positive
lexical_diversity
emotion_sadness
certainty_x_caps
word_count_x_lexical_div
rdoc_arousal_total
rdoc_arousal_dimension
hedge_x_exclaim
rdoc_reward_expectancy
rdoc_positive_valence_total
emotion_disgust
rdoc_valence_dimension
rdoc_threat_reward_ratio
rdoc_low_arousal
rdoc_arousal_net
sentiment_score
exclamation_count
word_count
rdoc_vader_compound
rdoc_emotional_intensity
sentiment_negative
emotion_joy
emotion_fear
rdoc_emotional_complexity
flesch_reading_ease
sentiment_neutral
emotion_surprise
rdoc_intensity_modifiers
rdoc_loss
conspiracy_markers
anecdote_x_exclaim
has_personal_anecdote

Table 3: Full SHAP feature importance rankings for seed oils and omega-6 domains (55 features per domain).

Ontological Validation of Biomedical Topic Models: SNOMED CT Hierarchy Distance as an Automated Evaluation Metric

Ilan S. Rubinfeld[1] **Sami Zaidi**[1] **Milosh Djuric**[1] **Mouhammad Halabi**[1]
Loay Kabbani[1] **Alex Shepard**[1] **Mohammad Ghassemi**[2]
[1]Henry Ford Health, Detroit, MI, USA
[2]Michigan State University, East Lansing, MI, USA
Correspondence: szaidi7@hfhs.org

Abstract

Standard coherence metrics for biomedical topic models encode no clinical knowledge and cannot detect clinically implausible topic groupings. We propose SNOMED CT Wu–Palmer hierarchy distance as a post hoc, ontology-grounded diagnostic. On vascular surgery (47,318 articles) and craniofacial surgery (27,493 articles) corpora, the metric flags clinically heterogeneous topics that coherence misses—e.g., abdominal aortic aneurysm repair grouped with deep vein thrombosis ($d = 0.600$). Diagnostic signals are nearly identical across eight BERTopic embedding strategies including ontology-enhanced models, but diverge across model families: BERTopic alone produces a positive within- vs. cross-topic Cohen's d, while LDA, NMF, and Top2Vec at matched topic counts score below their own cross-topic baselines (Cohen's $d < 0$; Mann–Whitney $p > 0.99$). The score is therefore sensitive to topic-model output choice, not only to embedding choice within a single pipeline. A pre-clustering screening experiment finds near-zero correlation ($|\rho| < 0.08$) between embedding cosine and SNOMED CT similarity, arguing that ontological validation belongs after clustering rather than as an embedding screen. We additionally describe a two-stage UMLS-CUI stopword filter that preserves high-frequency domain-specific concepts which naive frequency filtering would discard. After one-time concept curation, the diagnostic itself is automated and requires no per-topic expert scoring.

1 Introduction

Topic models have become a standard tool for organizing and exploring biomedical literature at scale (Yeganova et al., 2018). Latent Dirichlet Allocation (Blei et al., 2003) and, more recently, BERTopic (Grootendorst, 2022) are widely applied in bibliometric analyses across clinical specialties, enabling researchers to identify thematic clusters, track temporal trends, and map the intellectual landscape of a field.

Evaluation of topic model quality relies primarily on intrinsic coherence metrics such as u_{mass} and C_v, which measure word co-occurrence statistics within discovered topics (Röder et al., 2015). While these metrics capture statistical regularity, they encode no domain knowledge. As a concrete example, a topic that groups articles on abdominal aortic aneurysm (AAA) repair—an elective *arterial reconstructive procedure*—with articles on deep vein thrombosis (DVT)—a *venous thromboembolic disease*—may achieve acceptable coherence scores because both share vascular vocabulary ("treatment," "graft," "outcome"). Yet the two occupy distinct branches of every clinical taxonomy, and merging them would mislead any researcher relying on the topic structure.

This gap between statistical coherence and clinical coherence has practical consequences. Researchers using topic models for systematic review screening (Thomas et al., 2011), research gap identification, or funding landscape analysis may draw conclusions from topics that conflate clinically distinct entities. A researcher asking "how has vascular surgery research evolved?" using a model that merges AAA repair and DVT into one topic is analyzing an artifact, not a literature structure.

The alternative—human expert evaluation—does not scale. Chang et al. (2009) demonstrated that human topic interpretability judgments are subjective and poorly correlated with statistical metrics. Hoyle et al. (2021) showed that even expert coherence ratings vary substantially, concluding that human evaluations have been largely abandoned by topic model developers. For a vascular surgery corpus producing 37 topics across 40 mapped clinical concepts, exhaustive pairwise clinical review requires evaluating $\binom{40}{2} = 780$ concept pairs—an effort no clinical reviewer can perform systematically, and one where reviewer differences

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 342–352
July 3-4, 2026 ©2026 Association for Computational Linguistics

in training, subspecialty, and clinical culture would introduce arbitrary variation (Lau et al., 2014). An automated, ontology-grounded method is needed.

We address this gap by introducing SNOMED CT hierarchy distance as a post hoc, ontology-grounded diagnostic for biomedical topic models. SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms) is the most comprehensive clinical terminology, organizing over 350,000 active concepts in a directed acyclic graph via `|is-a|` relationships (Pedersen et al., 2007; Melton et al., 2006). We apply Wu–Palmer normalized distance (Wu and Palmer, 1994) based on the Lowest Common Ancestor (LCA) to score the ontological coherence of topics produced by BERTopic-style pipelines (neural embeddings followed by UMAP+HDBSCAN clustering), and additionally compare against LDA, NMF, and Top2Vec under matched topic counts to test whether the diagnostic generalizes across topic-model families (§4.6). Contextualized topic models remain future work.

Our contributions are:

1. **Ontology-grounded post hoc diagnostic for biomedical topic models.** We define an SNOMED CT Wu–Palmer distance metric over topic-model outputs and show, on vascular (47,318 articles) and craniofacial (27,493 articles) corpora, that it flags clinically heterogeneous topics that coherence scores do not.

2. **Sensitivity across topic-model families.** The score produces near-identical signals across eight BERTopic embedding strategies (mxbai, PubMedBERT, SapBERT, BioLORD, cui2vec, MeSH TF-IDF, CUI co-occurrence, and an ensemble) but diverges across model families: BERTopic produces the only positive within- vs cross-topic Cohen's d on both corpora, while LDA, NMF, and Top2Vec at matched topic counts all score below their own cross-topic baselines (Cohen's $d < 0$, Mann–Whitney $p > 0.99$). This shows the metric is sensitive to topic-model output choice and not only to embedding choice within a single pipeline.

3. **Negative embedding screening result.** Across ten model–corpus combinations the cosine geometry of biomedical embeddings is only weakly correlated with SNOMED CT distance (Spearman $|\rho| < 0.08$), arguing that ontological validation is best applied after clustering rather than as an embedding screen—even for models trained on UMLS relations.

4. **Two-stage CUI stopword filtering.** We identify generic UMLS concepts (e.g., *Patients*, *Clinical Research*) that inflate CUI-based topic overlap scores and cascade-merge BERTopic outputs, and propose a frequency-plus-semantic-type filter that preserves high-frequency domain-specific concepts (e.g., *Cleft Palate* 19.1%, *Complication* 22.5%) which naive frequency filtering would incorrectly drop.

2 Background and Related Work

SNOMED CT and UMLS. SNOMED CT (Donnelly, 2006) is a clinical terminology of ∼350,000 active concepts organized as a polyhierarchical DAG via `|is-a|` relations; each concept has a stable SNOMED Concept Identifier (SCTID). The Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) aggregates over 200 vocabularies, assigns each concept a Concept Unique Identifier (CUI), and provides semantic types (e.g., T046 Pathologic Function) and crosswalks to MeSH, ICD, and SNOMED CT. In this paper *SCTID* refers to nodes used to compute hierarchy distance, and *CUI* refers to concepts extracted from text via SciSpacy (Neumann et al., 2019).

Topic models for biomedical literature. LDA (Blei et al., 2003) and its biomedical applications (Yeganova et al., 2018; Thomas et al., 2011) preceded the contemporary embedding-plus-clustering family typified by BERTopic (Grootendorst, 2022): encode with a language model, reduce via UMAP (McInnes et al., 2018), cluster with HDBSCAN. Domain-pretrained encoders such as PubMedBERT (Gu et al., 2021), SapBERT (Liu et al., 2021), BioLORD (Remy et al., 2024), and CUI-aware spaces such as cui2vec (Beam et al., 2020) are widely substituted at the embedding stage. Our primary experiments use this family; we additionally compare against LDA, NMF, and Top2Vec (§4.6) to test whether the diagnostic generalizes beyond embedding-based clustering. Contextualized topic models remain future work (§7). Topic quality is typically assessed via word-co-

occurrence coherence (Röder et al., 2015), which Chang et al. (2009) and Hoyle et al. (2021) show agrees poorly with human judgement, motivating automated, knowledge-grounded alternatives (Lau et al., 2014).

Ontology-grounded similarity. A long line of work uses hierarchical taxonomies to score concept similarity—Wu–Palmer (Wu and Palmer, 1994), Resnik, Lin, Jiang–Conrath—surveyed in Harispe et al. (2015) and applied to biomedical ontologies in Pedersen et al. (2007). SNOMED CT distance has been used for patient similarity (Melton et al., 2006). Bhattacharya et al. (2018) applied topic modeling *to* SNOMED codes in EHR data; we work in the opposite direction. The distinction between distributional *relatedness* and taxonomic *similarity* is established in Hill et al. (2015), and Mao and Fung (2020) showed that UMLS relation types are not uniformly captured by biomedical embeddings; our negative screening result (§4.4) extends these findings to SNOMED CT hierarchy distance specifically. To our knowledge, the present work is the first to apply SNOMED CT distance as a post hoc diagnostic on topic-model output of biomedical literature.

3 Methods

3.1 SNOMED CT Hierarchy Distance

We score topic-level ontological coherence using Wu–Palmer normalized distance (Wu and Palmer, 1994) over the SNOMED CT `|is-a|` hierarchy. Among the canonical taxonomy-based measures (Harispe et al., 2015), Wu–Palmer is depth-normalized (yielding $d \in [0, 1]$ for direct comparison across SNOMED branches of differing depth), does not require corpus-derived information content (unlike Resnik, Lin, and Jiang–Conrath, which would introduce a second corpus-dependent quantity into a metric meant to be *purely* ontological), and is the form most commonly applied to SNOMED CT in prior work (Pedersen et al., 2007).

For two concepts c_1 and c_2 with Lowest Common Ancestor (LCA) in the `|is-a|` hierarchy, Wu–Palmer normalized distance is

$$d(c_1, c_2) = 1 - \frac{2\,\mathrm{depth}(\mathrm{LCA})}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)},$$

where depth is the shortest path from the SNOMED CT root (Pedersen et al., 2007). The distance is 0 for identical concepts and approaches 1 for pairs sharing only a high-level ancestor. For polyhierarchical concepts with multiple `|is-a|` parents we use the shortest path, following standard practice (Harispe et al., 2015). We chose SNOMED CT over MeSH (another common UMLS-linked taxonomy) because SNOMED CT provides finer-grained subspecialty distinctions for surgical concepts and the procedural–disorder branch split that drives several of our representative failure cases; MeSH descriptors are nonetheless used in the embedding-comparison arm (Table 2, MeSH TF-IDF).

3.2 Concept-to-Article Mapping

For each specialty, we curate SNOMED CT concepts representing the major clinical domains: 75 concepts for vascular surgery (spanning aortic, carotid, peripheral arterial, venous, and access procedures and their corresponding disorders) and 48 concepts for craniofacial surgery (spanning congenital anomalies, facial trauma, reconstruction, tumors, and nerve injury). The asymmetry reflects vascular surgery's broader anatomical scope and broader publication remit: beyond procedure and disorder concepts spanning each subspecialty, the vascular set also covers pharmacotherapy, imaging, basic-science, and professional topics that the craniofacial registry, anchored more tightly to anatomy and pathology, largely omits. Each concept is associated with discriminating keywords (e.g., "carotid endarterectomy" → SCTID 174776006). An article maps to a concept by keyword search over both its title and abstract text—abstracts having been shown to carry the highest information density of any article section (Schuemie et al., 2004). Up to 100 articles per concept are retained.

Topic assignments from BERTopic provide the link between articles and topics, enabling determination of which SNOMED CT concepts are represented within each topic.

Coverage. Keyword-based mapping against both title and abstract text achieves 78.8% coverage of vascular articles ($n = 37{,}306$ of 47,318) and 61.0% of craniofacial articles ($n = 16{,}763$ of 27,493). All 37 vascular topics (100%) and 12 of 29 craniofacial topics (41.4%) contain $\geq$2 mapped concepts and are therefore evaluable. The remaining unmapped articles lack concept-specific keywords across both title and abstract text. Prior work found information density highest in abstracts, although full text provides broader coverage (Schuemie et al., 2004). MeSH-to-SNOMED crosswalks (Bodenrei-

der, 2004) could further improve coverage.

3.3 Topic-Level Ontological Validation

For each topic t containing mapped concepts $C_t = \{c_1, c_2, \ldots, c_k\}$ where $k \geq 2$, we compute the mean pairwise ontological distance:

$$\bar{d}(t) = \frac{2}{k(k-1)} \sum_{i<j} d(c_i, c_j) \qquad (1)$$

Using a heuristic default threshold of $\bar{d}(t) > 0.5$, topics are flagged as *clinically heterogeneous*. We additionally identify *illustrative cross-domain pairs*—specific concept pairs within a topic that make the heterogeneity clinically interpretable.

3.4 Embedding Screening (Stage 0)

We test whether embedding geometry mirrors clinical ontology, which would allow embedding selection to serve as an upstream proxy. For each model, we compute centroid embeddings per SNOMED CT concept by averaging embeddings of mapped articles, compute pairwise cosine similarities, and correlate with SNOMED CT Wu–Palmer similarity using Spearman's ρ.

3.5 CUI Stopwords for Topic Reconciliation

When using UMLS Concept Unique Identifiers (CUIs) extracted from title and abstract text via SciSpacy (Neumann et al., 2019) to compare topics—e.g., via Jaccard overlap of per-topic CUI profiles—generic medical concepts that appear across most articles inflate similarity scores and cause spurious cascade merges. We identify these "medical CUI stopwords" empirically via a two-stage procedure. First, for each CUI extracted from the corpus, we compute document frequency (the fraction of articles containing the concept). CUIs exceeding a 15% prevalence threshold are candidates for filtering. Second, we classify each candidate by its UMLS semantic type: CUIs belonging to generic types (T079 Temporal Concept, T078 Idea or Concept, T080 Qualitative Concept, T081 Quantitative Concept, T101 Patient or Disabled Group, T062 Research Activity, etc.) are designated stopwords, while CUIs belonging to clinical types (T046 Pathologic Function, T061 Therapeutic Procedure, T047 Disease or Syndrome, T023 Body Part, etc.) are retained regardless of frequency. This two-stage approach prevents filtering of high-frequency domain-specific concepts that are substantively important within a specialty corpus.

On the vascular corpus (47,318 articles), the frequency threshold identifies 27 CUIs at $\geq$15%, of which 21 are generic (led by *Patients*, C0030705, 77.7%) and 6 are clinical (e.g., *Complication*, 22.5%; *Interventional procedure*, 18.8%). The semantic type filter retains the clinical CUIs and removes only the 21 generic stopwords. Without filtering, CUI-based topic overlap scores are dominated by ubiquitous concepts; in early experiments, Jaccard-based merge rules cascaded 37 topics down to 2. After filtering, no spurious merges occur (Section 4.2).

3.6 Post-Modeling Ontology Reconciliation

As a complementary step, we apply a two-pass pipeline. After BERTopic clustering (Pass 1), a CUI-based reconciliation pass (Pass 2) scores all topic pairs on Jaccard overlap of their CUI profiles (with stopwords removed) and merges pairs exceeding an empirically determined similarity threshold of 0.7 (requiring substantial CUI overlap before merging), with an anti-cascade constraint (each topic merges at most once per pass). This tests whether statistically discovered topics contain redundant clinical content that should be consolidated.

3.7 Corpora

Vascular surgery. 47,318 articles from 8 vascular surgery journals,[1] spanning 1984–2026. Journals were identified via PubMed's NLM Catalog under the "Vascular Surgery" and "Vascular Diseases" Broad Subject Terms, selecting all journals whose primary scope is vascular surgical practice. To test metric robustness across upstream modeling choices, we apply SNOMED CT validation to topic model outputs from eight embedding strategies, selected to span the major design axes available for biomedical text representation: general-purpose transformer (mxbai-embed-large-v1, 335M parameters (Lee et al., 2024)), biomedical transformers (PubMedBERT, 110M (Gu et al., 2021); SapBERT, 110M (Liu et al., 2021)), ontology-grounded models (BioLORD-2023, 110M (Remy et al., 2024); cui2vec, 500-d (Beam et al., 2020); Poincaré embeddings of SNOMED CT `|is-a|` edges, 200-d), a CUI co-occurrence TF-IDF baseline (200-d via SVD), and an ensemble (L2-normalized average of PubMedBERT + BioLORD). All use

[1] JVS, Annals of Vascular Surgery, EJVES, J. Endovascular Therapy, Vascular, JVS-VLD, Seminars in Vascular Surgery, JVS-Vascular Science.

Model	All-pairs	Same-dom.	Weighted	Single.
mxbai	0.711	0.394	0.711	10
PubMedBERT	0.722	0.367	0.742	1
BioLORD	0.715	**0.337**	0.735	2
Ensemble	**0.701**	0.347	**0.711**	**0**

Table 1: SNOMED CT ontological validation applied to four embedding models on JVS ($n = 14{,}290$). The metric yields similar diagnostic signals ($\bar{d} \approx 0.70$–0.74) across models, while still revealing model-specific failure modes (e.g., mxbai produces 10 singleton concepts vs. 0 for ensemble).

BERTopic with UMAP + HDBSCAN clustering (McInnes et al., 2018) (`min_cluster_size` = 300, `min_samples` = 15).

Craniofacial surgery. 27,493 articles from 9 journals,[2] spanning 1994–2025, identified via PubMed NLM Catalog under craniofacial and oral surgery subject headings. BERTopic with mxbai-embed-large-v1 embeddings, producing 29 topics.

3.8 Classical Topic-Model Baselines

To test whether the SNOMED CT diagnostic discriminates across topic-model families (not only embedding choices within BERTopic), we additionally fit three classical models on both corpora with matching topic counts ($k = 37$ vascular, $k = 29$ craniofacial): LDA (Blei et al., 2003) over a bag-of-words representation (min_df = 5, max_df = 0.95), NMF over TF-IDF features (default sklearn settings), and Top2Vec, which jointly learns document and word embeddings followed by HDBSCAN clustering and hierarchical reduction to the target k. The same SNOMED concept mapping and Wu–Palmer scoring pipeline is then applied to each model's topic assignments.

4 Results

4.1 Vascular Surgery Corpus

To assess whether the metric produces stable diagnostic signals across different upstream embedding choices, we apply it to four embedding models on the single-journal JVS corpus (Table 1; 14,290 articles, 75 SNOMED concepts) and then extend to eight embedding strategies on the full multi-journal corpus (47,318 articles, 8 journals).

With under-constrained clustering (`min_cluster_size` = 40), the mxbai model produces a mega-cluster (Topic 0, 36% of corpus) spanning 4+ SNOMED branches with mean intra-topic distance = 0.768 across 78 concept pairs. Six illustrative cross-domain pairs show clinically implausible groupings within this single topic:

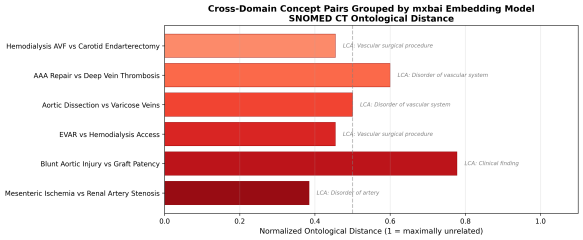


Figure 1: Six representative cross-domain concept pairs grouped within a single topic by the mxbai model. Each bar shows the Wu–Palmer SNOMED CT distance between clinically unrelated concepts that the topic model erroneously merged. Standard coherence metrics do not flag this topic.

- Hemodialysis AVF ↔ Carotid endarterectomy ($d = 0.455$; LCA: Vascular surgical procedure)
- AAA repair ↔ Deep vein thrombosis ($d = 0.600$; LCA: Disorder of vascular system)
- Aortic dissection ↔ Varicose veins ($d = 0.500$; LCA: Disorder of vascular system)
- Blunt aortic injury ↔ Graft patency ($d = 0.778$; LCA: Clinical finding)
- EVAR ↔ Saphenous vein ablation ($d = 0.500$; LCA: Endovascular procedure)
- PAD claudication ↔ IVC filter ($d = 0.714$; LCA: Disorder)

In this setting, the metric flags failures that coherence scores do not: the mega-cluster achieves acceptable coherence despite grouping concepts from 4+ SNOMED branches. Across all four models, the metric flags clinically heterogeneous topics (all-pairs $\bar{d}$ range: 0.701–0.722), suggesting that the diagnostic signal is not confined to a single embedding choice.

Multi-journal robustness (8 embedding strategies). On the full 47,318-article, 8-journal vascular corpus with appropriately constrained clustering (`min_cluster_size` = 300), we test eight embedding strategies (Table 2). Topic counts range

[2] J. Craniofacial Surgery, J. OMS, Cleft Palate–Craniofacial J., Int. J. OMS, J. CMF Surgery, PRS, Br. J. OMS, Oral Surg. Oral Med. Oral Pathol., Head & Face Medicine.

Embedding	K	Out.%	Sil.	C_v
Ensemble	**37**	**18.3**	0.526	**0.792**
SapBERT	33	23.9	0.465	0.757
SapBERT+Snomed2Vec	33	23.9	0.464	0.757
CUI co-occurrence	49	28.9	0.528	0.718
MeSH TF-IDF	37	22.3	0.597	0.639
SapBERT+cui2vec	20	26.6	0.391	0.796
cui2vec	5	6.7	0.170	0.392
Snomed2Vec	26	9.3	−0.065	—

Table 2: Eight embedding strategies on the vascular corpus ($n = 47{,}318$, 8 journals). K = topics, Out.% = outlier rate, Sil. = silhouette. Ontology-enhanced embeddings (Snomed2Vec, cui2vec) do not outperform the baseline ensemble; Snomed2Vec produces negative silhouette, indicating cluster overlap. Bold = recommended model.

CUI (Name)	Doc Freq	%
C0030705 (Patients)	36,752	77.7
C0008972 (Clinical Research)	18,602	39.3
C0439234 (year)	16,739	35.4
C0868928 (Case)	12,990	27.5
C0039798 (therapeutic)	12,360	26.1
C0332281 (Associated with)	12,278	25.9
C0220825 (Evaluation)	11,707	24.7
…	(27 total $\geq 15\%$)	

Table 3: Top generic UMLS CUIs by document frequency in the vascular corpus. These "medical stopwords" must be filtered before CUI-based topic comparison to prevent cascade merges.

from 5 (cui2vec, severe under-clustering) to 49 (CUI co-occurrence). The recommended model (PubMedBERT + BioLORD ensemble, 37 topics, $C_v = 0.792$, 18.3% outliers) eliminates the mega-cluster problem observed with under-constrained parameters, producing topics that domain experts can inspect. Across the six viable arms, mean within-topic distances range from 0.767 to 0.770 (cross-topic baseline: 0.771), with Cohen's d between 0.003 and 0.015. This consistency suggests the findings are not artifacts of a single embedding model. The small positive Cohen's d reflects vascular topics spanning both procedures and disorders within each domain.

Clustering parameters vs. embedding choice. The jump from `min_cluster_size` $= 40$ (43 topics, 30.2% outliers, mega-clusters) to `min_cluster_size` $= 300$ (37 topics, 18.3% outliers, no mega-clusters) suggests that HDBSCAN parameter tuning has a larger effect on clinical face validity than embedding model selection. A topic count in the 30–50 range provided the clearest balance of subspecialty granularity in our data, while under-constrained clustering ($K > 100$) or over-constrained clustering ($K < 20$) produced topics that were implausibly coarse or fragmented.

4.2 CUI Stopwords

Table 3 shows the ten most prevalent CUIs in the vascular corpus. Without filtering, CUI Jaccard overlap between topic pairs is dominated by these generic concepts: in early experiments without a stoplist, Pass 2 reconciliation cascade-merged 37 topics to 2. After filtering the 21 generic CUIs (retaining 6 clinical CUIs via semantic type classification), Pass 2 produces zero merges in our experiments across all eight embedding arms—suggesting that appropriately constrained HDBSCAN already produces non-redundant topics.

Cross-corpus validation. To test whether the same generic-term pattern appears in a second corpus, we performed an analogous analysis on the craniofacial corpus using MeSH descriptor document frequencies (27,305 articles). Of 12 MeSH terms at $\geq$15%, 9 are generic (Humans 93.4%, Male 58.7%, Female 57.5%, Adult 36.7%, Retrospective Studies 25.4%) and 3 are clinical (Mandible 19.3%, Cleft Palate 19.1%, Cleft Lip 16.4%). Frequency-only filtering would incorrectly remove *Cleft Palate*—a defining condition of the specialty appearing in one-fifth of articles—while semantic type classification correctly retains it. The same structural pattern (generic demographics and study-design terms dominating, with domain-specific clinical terms at the boundary) appears in both corpora, supporting the two-stage approach as reusable across these two specialties.

4.3 Craniofacial Surgery Corpus

Table 4 presents per-topic ontological distances for the 12 craniofacial topics containing $\geq$2 mapped SNOMED concepts (48 concepts, 61 intra-topic pairs).

Nine topics ($\bar{d} < 0.5$) are ontologically coherent. T8 (Facial trauma) groups mandible, orbital, zygoma, nasal, and Le Fort fractures—all under "Fracture of bone of face" in SNOMED CT ($\bar{d} = 0.136$). T3 (Cleft lip and palate) groups cleft lip, cleft palate, combined cleft, and Pierre Robin sequence ($\bar{d} = 0.220$).

Three topics ($\bar{d} \geq 0.5$) are flagged as heterogeneous. Representative examples include:

- Rhinoplasty $\leftrightarrow$ Mandible fracture ($d = 1.000$;

Topic	$\bar{d}$	k	Pairs
T2: Orthognathic surgery	0.116	3	3
T8: Facial trauma	0.136	5	10
T1: Bone / distraction	0.200	2	1
T3: Cleft lip and palate	0.220	4	6
T7: Nerve injury	0.278	3	3
T13: Fillers / fat grafting	0.333	2	1
T20: Beauty / aesthetics	0.333	2	1
T6: Tumors and cysts	0.359	3	3
T0: Craniosynostosis	0.434	8	28
T12: Condylar / TMJ	0.511	3	3
T4: Flap reconstruction	0.600	2	1
T16: Hemangioma / vasc. malf.	0.600	2	1

Table 4: Per-topic SNOMED CT distances for craniofacial surgery ($n = 27{,}493$, mxbai embeddings). The midrule separates topics below the heuristic threshold ($\bar{d} < 0.5$) from those at or above it ($\bar{d} \geq 0.5$). k = number of mapped SNOMED concepts.

LCA: SNOMED CT root — elective aesthetic procedure vs. acute trauma)

- Cleft lip ↔ Ameloblastoma ($d = 0.667$; LCA: Disease — congenital anomaly vs. neoplasm)
- Craniosynostosis ↔ Burn injury ($d = 0.556$; LCA: Disease — congenital vs. traumatic)
- Facial nerve palsy ↔ Osteonecrosis of jaw ($d = 0.636$; LCA: Disease — neurological vs. bone pathology)

This thresholding provides a simple operational mechanism for topic-level filtering, enabling identification of topics that may require splitting or further inspection. The overall mean of per-topic mean distances is 0.343 (SD = 0.159), compared to 0.768 for the mxbai vascular mega-cluster. The craniofacial corpus shows better topic-level ontological coherence overall, likely because the 29-topic structure provides finer granularity than the mxbai vascular run's coarser clustering.

Threshold sensitivity. We adopt $\bar{d} > 0.5$ as a heuristic default separating single-domain topics from cross-domain groupings. Discriminative power varies by specialty and granularity—vascular topics inherently bridge procedure and disorder branches—and specialty-specific calibration is recommended.

4.4 Embedding Screening: A Negative Result

Table 5 presents Spearman correlations between embedding cosine similarity and SNOMED CT Wu–Palmer similarity.

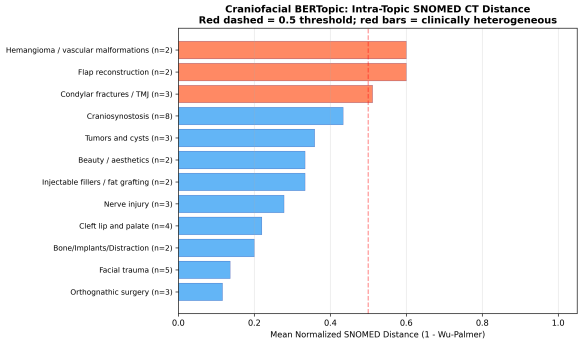


Figure 2: Per-topic SNOMED CT ontological distances for the craniofacial corpus. Topics below the dashed threshold ($\bar{d} = 0.5$) are ontologically coherent; topics above are flagged as clinically heterogeneous.

Model	Corpus	ρ	p	Pairs
mxbai	Vasc. (JVS)	0.037	0.305	780
PubMedBERT	Vasc. (JVS)	0.065	0.071	780
BioLORD	Vasc. (JVS)	0.013	0.713	780
Ensemble	Vasc. (JVS)	0.045	0.211	780
Ensemble	Vasc. (8J)	0.005	0.900	780
SapBERT	Vasc. (8J)	0.004	0.912	780
CUI co-occ.	Vasc. (8J)	0.023	0.522	780
MeSH TF-IDF	Vasc. (8J)	0.043	0.234	780
SapBERT+cui2vec	Vasc. (8J)	0.025	0.493	780
mxbai	Craniofacial	0.022	0.468	1,128

Table 5: Embedding screening: Spearman ρ between cosine similarity and SNOMED CT Wu–Palmer similarity across 10 model–corpus combinations. All show $|\rho| < 0.08$. Vasc. (JVS) = single-journal; Vasc. (8J) = 8-journal corpus.

All ten model–corpus combinations show near-zero correlation ($|\rho| < 0.08$). Most comparisons come from the vascular corpus, but the same near-zero pattern also appears in the craniofacial mxbai run. Within vascular data, the result is consistent across general-purpose (mxbai), biomedical (PubMedBERT, SapBERT), ontology-grounded (BioLORD, CUI co-occurrence, MeSH TF-IDF), and hybrid models, and across both single-journal and multi-journal settings. Notably, BioLORD-2023—which was explicitly trained on UMLS concept relationships including SNOMED CT (Remy et al., 2024)—shows $\rho = 0.013$ ($p = 0.71$).

Embedding models organize biomedical text by literature co-occurrence patterns, not by clinical ontological relationships (Hill et al., 2015). Concepts that are ontologically distant (e.g., AAA and DVT: different vascular systems, different pathophysiology) co-occur in the same journals, the same patient populations, and the same review articles. The embedding space reflects this co-occurrence

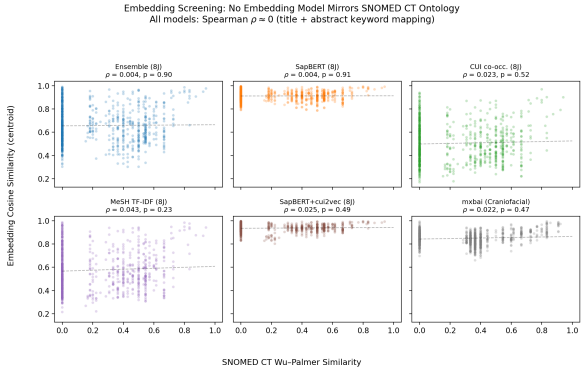


Figure 3: Embedding screening: cosine similarity versus SNOMED CT Wu–Palmer similarity for all concept pairs. All ten model–corpus combinations show Spearman $\rho \approx 0$, suggesting that embedding geometry does not mirror clinical ontological structure.

structure—distributional *relatedness*—not the clinical taxonomy—ontological *similarity* (Hill et al., 2015).

4.5 Informal Clinician Face-Validity Check

We include an informal three-surgeon assessment of one pipeline arm to illustrate—not to validate—the diagnostic's qualitative behaviour. Two vascular surgeons and one acute care surgeon *jointly* discussed the 37 topics from Arm A (ensemble) and proposed merge and split groupings; ratings were not collected independently, no blinding was applied, and no formal scoring rubric was used. We then compared this shared assessment with a domain-level split audit that classifies each topic by its distribution across nine vascular anatomical domains (Aortic, Cerebrovascular, Peripheral, Venous, Access, Visceral, Renal, Trauma, Thoracic Outlet). This comparison only covers Arm A on the vascular corpus, 16 judgements, and one specialty; it should not be read as evidence that the metric tracks clinical judgement in general (§7).

On these 16 judgements, the metric agreed with the surgeons on 4 merge groups comprising 13 topics in 12/13 cases (92%; e.g., Topics 7, 15, 23, 35—aortic aneurysm repair—scored zero heterogeneity with >83% Aortic concentration) and on 2 of 3 split requests (67%); the one miss separated amputation from other peripheral arterial topics, an intra-domain distinction that domain-level analysis cannot capture. Combined directional agreement was 14/16 (88%). We report this number for context only; it cannot establish metric validity. A blinded, independently rated, multi-specialty clinician study with inter-rater agreement is needed and is left to future work.

Corpus	Method	Eval.	$\bar{d}_w$	$\bar{d}_c$	Cohen's d
Cranio.	BERTopic	12/29	0.343	0.768	+0.159
	LDA	28/29	0.715	0.703	−0.043
	NMF	29/29	0.726	0.713	−0.046
	Top2Vec	29/29	0.726	0.712	−0.050
Vasc.	BERTopic	37/37	0.767	0.771	+0.015
	LDA	36/37	0.767	0.747	−0.073
	NMF	37/37	0.770	0.750	−0.071
	Top2Vec	37/37	0.769	0.749	−0.071

Table 6: SNOMED diagnostic applied to classical and neural topic models. $\bar{d}_w$ = mean within-topic distance; $\bar{d}_c$ = cross-topic baseline; Eval. = topics with $\geq$2 mapped concepts. Positive Cohen's d indicates within-topic < cross-topic (clinically coherent grouping). All three alternatives show negative Cohen's d on both corpora (Mann–Whitney $p > 0.99$).

4.6 Classical Baseline Comparison

To test whether the diagnostic generalizes beyond BERTopic-style pipelines, we apply it to three alternative topic models—Latent Dirichlet Allocation (LDA; Blei et al. 2003), Non-negative Matrix Factorization (NMF), and Top2Vec—run on both corpora with matching topic counts ($k = 37$ vascular, $k = 29$ craniofacial). LDA uses bag-of-words input, NMF uses TF-IDF, and Top2Vec jointly learns document and word embeddings with hierarchical topic reduction.

Table 6 shows that all three classical alternatives score below their own cross-topic baselines on both corpora, while BERTopic alone produces positive within- vs. cross-topic separation. On the craniofacial corpus, BERTopic achieves $\bar{d}_w = 0.343$ (well below the cross-topic baseline of 0.768), while LDA, NMF, and Top2Vec produce $\bar{d}_w \approx 0.72$—statistically indistinguishable from their own cross-topic baselines (Mann–Whitney $p > 0.99$; Cohen's $d < 0$). LDA generates 10 mega-clusters containing >40 mapped concepts each (of 48 total), indicating that most concepts co-occur in nearly every topic; NMF and Top2Vec produce 6 and 3 mega-clusters respectively. On the vascular corpus, all four methods show high $\bar{d}_w \approx 0.77$, but only BERTopic produces a (small) positive Cohen's d.

5 Discussion

What the metric does, and what it does not do. The diagnostic flags topics whose mapped SNOMED CT concepts span distant hierarchy branches—a failure mode that coherence metrics, which score word co-occurrence rather than on-

tological position, do not detect. We frame the metric as a post hoc complement to coherence, not a replacement: a topic may be coherent yet ontologically heterogeneous, and vice versa. After one-time concept curation, the scoring step is automated, but curation, threshold selection, and the interpretation of high-distance topics still require domain judgement.

The co-occurrence–ontology gap. The near-zero screening correlation ($|\rho| < 0.08$ across ten model–corpus combinations) extends Mao and Fung (2020) to the specific case of SNOMED CT hierarchy distance: even models trained on UMLS relations (BioLORD-2023, SapBERT) produce embedding spaces that capture distributional relatedness, not taxonomic position. The practical implication is that embedding choice cannot stand in for ontological validation, and that ontological scoring is best applied after clustering.

Clustering constraints matter more than embedding choice. Our eight-arm comparison suggests HDBSCAN parameter selection matters more for clinical face validity than embedding choice within the BERTopic family. Under-constrained clustering (`min_cluster_size` $= 40$) produces mega-clusters spanning 4+ SNOMED branches, whereas over-constrained settings ($K < 20$) collapse subspecialty distinctions. In our data, 30–50 topics gave the best balance, so future studies should report HDBSCAN sensitivity alongside embedding selection.

Sensitivity across topic-model families. The diagnostic is a post hoc score, not a constraint on the underlying models. What the cross-family comparison shows is that the score is sensitive to topic-model output choice, not only to embedding choice within a single pipeline (Table 2 vs. Table 6). We do not claim that BERTopic is universally superior to LDA, NMF, or Top2Vec; we observe only that its outputs have higher SNOMED CT ontological coherence on these two corpora at the tested topic counts. The cross-family direction is consistent across both corpora, but the magnitude of BERTopic's advantage is much larger on craniofacial ($+0.159$) than on vascular ($+0.015$).

Medical CUI stopwords. Generic UMLS concepts dominate CUI-based topic overlap scores in our data; any system using CUI profiles for clustering or merge decisions will encounter the same noise floor. The two-stage filter (frequency $\geq$15% followed by UMLS semantic type) preserves high-frequency domain-specific terms (*Cleft Palate* 19.1%, *Complication* 22.5%) that naive frequency thresholding would discard, while removing demographic, study-design, and temporal qualifiers. The generic/substantive boundary remains context-dependent; semantic types provide a principled default with room for domain-specific override. Pass 2 ontology reconciliation produced zero merges across all eight arms, indicating that constrained HDBSCAN already discovers non-redundant clinical topics in our data.

6 Conclusion

SNOMED CT Wu–Palmer distance flags clinically heterogeneous biomedical topics that word-coherence metrics do not, and it is sensitive to topic-model output choice across families. Across two surgical corpora, eight BERTopic embedding strategies, and the LDA, NMF, and Top2Vec baselines, only BERTopic produces positive within- vs. cross-topic separation (Tables 2, 6). The near-zero embedding-screening correlation argues that ontological validation belongs *after* clustering, not as an embedding screen. We additionally contribute a two-stage CUI stopword filter and find that HDBSCAN constraints may matter more for clinical face validity than embedding choice. Contextualized topic models and a blinded independently rated clinician study remain future work (§7).

7 Limitations

Scope of topic-model family. The primary experiments use BERTopic with UMAP+HDBSCAN. A preliminary comparison with LDA, NMF, and Top2Vec (§4.6) shows the diagnostic generalizes across topic-model families, but contextualized topic models, MeSH-guided variants, and systematic hyperparameter sweeps remain untested. The screening result (§4.4) specifically pertains to embedding-based pipelines.

Concept curation and coverage. Concept curation requires domain knowledge, though the effort is modest (∼50–75 concepts per specialty). Keyword mapping against title and abstract text covers 78.8% of vascular and 61.0% of craniofacial articles; full-text coverage would be broader (Schuemie et al., 2004), and MeSH-to-SNOMED crosswalks via UMLS (Bodenreider, 2004) could improve recall. Topics with fewer than two mapped concepts cannot be evaluated.

Metric calibration. SNOMED CT branch depths vary, and the $\bar{d} > 0.5$ flagging threshold is heuristic and specialty-sensitive; SNOMED CT also does not always capture cross-cutting themes (complications, perioperative care, imaging findings, outcomes) the way clinicians conceptualize them. Elevated distance should prompt inspection rather than be treated as categorical error. The CUI stoplist should be re-derived per specialty.

Method assumptions. Several methodological choices warrant flagging. The article-to-concept mapping caps each concept at 100 articles, which may distort topic statistics for high-frequency concepts; topics with only two mapped concepts are flagged on a single pair, exaggerating noise. The embedding-screening centroid aggregation may wash out fine-grained substructure, and the Spearman correlation assumes a monotonic alignment between embedding cosine and SNOMED CT similarity that need not hold. Wu–Palmer depth normalization assumes comparability across SNOMED branches, which may not hold uniformly between procedural and disorder hierarchies. Both validated corpora are surgical specialties; generalization to non-procedural domains (e.g., psychiatry, primary care, genomics) is untested, and temporal-span differences between corpora (1984–2026 vs. 1994–2025) are not analyzed.

Clinician validation. The three-surgeon assessment (§4.5) was a single joint discussion on one pipeline arm in one specialty, with no independent scoring, blinding, or rubric. It cannot establish metric validity. A blinded, independently rated, multi-specialty study with inter-rater agreement, power calculation, and a formal rubric is needed before the metric should be cited as predictive of clinical judgement.

References

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. *Pacific Symposium on Biocomputing*, 25:295–306.

Moumita Bhattacharya, Claudine Jurkovitz, and Hagit Shatkay. 2018. Co-occurrence of medical conditions: Exposing patterns through probabilistic topic modeling of SNOMED codes. *Journal of Biomedical Informatics*, 82:31–40.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22, pages 288–296.

Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. ACL.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread – new fluffy embeddings model. Mixedbread AI Blog. `https://www.mixedbread.ai/blog/mxbai-embed-large-v1`.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238. ACL.

Yuqing Mao and Kin Wah Fung. 2020. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *Journal of the American Medical Informatics Association*, 27(10):1538–1546.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29):861.

Genevieve B. Melton, Sonia Parsons, Frances P. Morrison, Adam S. Rothschild, Marianthi Markatou, and George Hripcsak. 2006. Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*, 39(6):697–705.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. ACL.

Ted Pedersen, Serguei V. S. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM.

Martijn J. Schuemie, Marc Weeber, Bob J. A. Schijvenaars, Erik M. van Mulligen, Cornelis C. van der Eijk, Rob Jelier, Barend Mons, and Jan A. Kors. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604.

James Thomas, John McNaught, and Sophia Ananiadou. 2011. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138. ACL.

Lana Yeganova, Sun Kim, Grigory Balasanov, and W. John Wilbur. 2018. Discovering themes in biomedical literature using a projection-based algorithm. *BMC Bioinformatics*, 19:269.

Systematic Evaluation of the Quality of Synthetic Clinical Notes Rephrased by LLMs at Million-Note Scale

Jinghui Liu[1] **Sarvesh Soni**[2] **Anthony Nguyen**[1]
[1]Australian e-Health Research Centre, CSIRO, Australia
[2]National Library of Medicine, National Institutes of Health, USA
jinghui.liu@csiro.au

Abstract

Large language models (LLMs) can generate or synthesize clinical text for a wide range of applications, from improving clinical documentation to augmenting clinical text analytics. Yet evaluations typically focus on a narrow aspect – such as similarity or utility comparisons – even though these aspects are complementary and best viewed in parallel. In this study, we aim to conduct a systematic evaluation of LLM-generated clinical text, which includes intrinsic, extrinsic, and factuality evaluations of synthetic clinical notes rephrased from MIMIC databases at million-note scale. Our analysis demonstrates that synthetic notes preserve core clinical information and predictive utility for coarse-grained tasks despite substantial linguistic changes, but lose fine-grained details for task like ICD coding. We show this loss of detail can be substantially mitigated by rephrasing notes by chunks rather than by the whole note, but at the cost of reduced factual precision under incomplete context. Through fact-checking and error analysis, we further find that synthesis errors are dominated by misinterpretation of clinical context, alongside temporal confusion, measurement errors, and fabricated claims. Finally, we show that the synthetic notes – despite their task-agnostic nature – can effectively augment task-specific training for rare ICD codes.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in generating human-like text, including text written by domain experts (Shao et al., 2024). In healthcare, studies have shown that LLMs can generate clinical notes similar to those drafted by clinicians (Van Veen et al., 2024; Small et al., 2025; Kahl et al., 2026), driving interest in LLM-assisted clinical note-taking to reduce the significant burden of documentation (Roberts, 2024; Colicchio et al., 2019). The flexible mechanism of generative LLMs also naturally lends them to synthetic data generation, where LLMs can be prompted to produce task-specific, synthetic clinical texts to help provide larger training samples for improving model performance and fairness in healthcare (Juwara et al., 2024; Ktena et al., 2024), especially since real-world, private patient records are often difficult to access.

Given the growing adoption of LLMs in clinical text generation, it is essential to measure the quality of the generated text. Studies on LLM-generated clinical documentation typically compare the generated notes with human-written reference notes via an array of textual similarity metrics (Croxford et al., 2025; Carandang et al., 2025), which measure the intrinsic quality of the LLM generation. For studies using LLMs for data synthesis, the evaluation is often performed based on a specific downstream task (Loni et al., 2025; Alshaikhdeeb et al., 2025), which focuses on the extrinsic utility of the generated content. While exceptions exist (Wang et al., 2024b), these different evaluation types are rarely presented jointly on a consistent set of synthetic clinical notes. Since existing metrics typically capture only a limited aspect of textual quality, a systematic evaluation combining multiple strategies, spanning multiple metrics and downstream tasks, is vital to understand how LLMs generate clinical text.

In this study, we aim to address this gap through a multi-angle evaluation of LLM-generated clinical notes, including 1) **similarity-based intrinsic evaluation**, 2) **utility-based extrinsic evaluation**, 3) **fairness evaluation in downstream applications**, 4) **fact-checking with LLM-as-judge**, and 5) **human analysis of LLM errors**. We apply these setups to clinical notes rephrased by LLMs, which we treat as synthetic notes.

Rephrasing can be viewed as the most basic form of text generation given that LLMs are provided with complete input contexts, and serves as a lean test case of whether LLMs can interpret clinical in-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 353–371
July 3-4, 2026 ©2026 Association for Computational Linguistics

formation in its original form. Furthermore, LLMs are considered particularly proficient in rephrasing noisy input into high-quality text, a technique now used widely in curating large-scale web corpora (Maini et al., 2025; Li et al., 2024). Intuitively, LLMs should be able to rephrase clinical notes with ease, but our findings reveal that significant information loss can occur, especially when rephrasing whole notes compared to smaller chunks — a finding demonstrated and cross-checked across our evaluation setups.

Additionally, we examine generation scenarios with varied prompts and repeated generation, and discuss their implications. Although the rephrased synthetic notes are task-agnostic, we show they can be valuable resources for data augmentation in a task-specific setting — demonstrated through ICD prediction (Ji et al., 2024; Gan et al., 2025). The benefits are particularly substantial for modeling rare ICD codes. These experiments shed light on the strengths and limitations of LLM-generated clinical notes.

2 Related Work

Evaluation is key to driving the development of models in NLP. In the clinical domain, proper evaluation is critical to ensure an NLP system is fit-for-purpose and can be applied in healthcare, where both accuracy and safety are critical. For LLM-generated text, existing works usually evaluate it by comparing to the human-written counterparts. For example, studies on generating clinical documentation using LLMs (Chung et al., 2025; Soni and Demner-Fushman, 2024) — such as drafting a discharge summary based on patient records (Schwieger et al., 2024) or writing a report from a medical image (Yu et al., 2023; Nicolson et al., 2024) — typically compare the generated notes with ground-truth notes written by human experts. This often involves multiple textual similarity metrics, such as ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020). Crafting pairwise similarity metrics is a research field in itself (Croxford et al., 2025).

Another application of LLM-based text generation is clinical text synthesis (Mitra et al., 2025; Liu et al., 2024a; Li et al., 2023), which provides data augmentation for a given task. These studies typically focus on particular note types or labeled datasets, and perform evaluation based on the downstream task of interest (Tang et al., 2023; Soni and Roberts, 2020), with task accuracy serving as the extrinsic measure. Unlike the aforementioned works, leveraging downstream tasks sidesteps the need for expert-written reference notes that are costly to produce at scale.

It is known that many existing metrics for text evaluation are imperfect (Sai et al., 2022) and only capture a specific aspect of textual features. Therefore, it is common to aggregate various metrics when comparing and ranking the performance of different generation models, and such aggregation tends to correlate better with human judgment (Xu et al., 2024; Van Veen et al., 2024). Furthermore, combining and analyzing different evaluation approaches can better illuminate generation quality by allowing results to be compared and cross-checked. In this study, we conduct such a systematic, multi-angle evaluation on a consistent set of synthetic notes generated at large scale — a combination that, to our knowledge, is still lacking in prior work.

Clinical text is noisy and varies in quality, a characteristic that resembles web data. Curating web data is an important research direction given its key role in training language models (Li et al., 2024), and this process can benefit from LLM-based synthesis. LLMs can be instructed to generate content with high "educational value" (Penedo et al., 2024; Gunasekar et al., 2023) or to produce improved versions of existing text snippets (Maini et al., 2024; Kang et al., 2025). The latter is also known as *rephrasing*. It has been found that even small-sized LLMs (<8B) are capable of producing high-quality rephrased corpora for improved and more efficient pretraining, and that larger LLMs do not necessarily improve downstream pretraining quality (Kang et al., 2025; Maini et al., 2024). This method has also been shown effective in the medical domain (Liu and Nguyen, 2024) for domain-specific pretraining compared to text synthesis from patient profiles alone (Kweon et al., 2024). Given its effectiveness, general applicability, and natural pairing with human-written notes for intrinsic evaluation, we adopt this method to generate synthetic clinical notes for our multi-angle evaluation.

3 Synthetic Clinical Notes by Rephrasing the MIMIC Databases

We generate synthetic clinical notes using LLMs by grounding on human-written notes, which are sourced from the MIMIC databases. We describe the synthesis process in the following.

3.1 Data Preparation

The MIMIC databases provide de-identified, real-world clinical data across various modalities. We considered all clinical notes from MIMIC-III (v1.4) (Johnson et al., 2016) and MIMIC-IV (v2.2) (Johnson et al., 2023). MIMIC-IV includes two note types (i.e., discharge summary and radiology report) while MIMIC-III provides a broader range of documentation charted during hospital encounters (e.g., physician and nursing notes). We align the format of the de-identification placeholders to underscores ("___"), consistent with MIMIC-IV. After deduplicating notes with the exact same content, the final pool of MIMIC notes comprises *4.6M clinical notes* with *1.4B words*.

3.2 Rephrasing Setup

Two approaches were considered in rephrasing clinical notes. The first is to rephrase *by the entire note*, preserving the original context. The second splits long notes into chunks and rephrases them *by chunk*. This is motivated by findings that LLMs may struggle to process information as the input context grows longer (Liu et al., 2024b; Hsieh et al., 2024), including during rephrasing (Maini et al., 2024). Each chunk contains approximately 150 words with sentence boundaries maintained, resulting in an average of 2.3 chunks per MIMIC note.

Small-scale generative LLMs are capable of effective rephrasing while maintaining efficiency (Kang et al., 2025; Maini et al., 2024). Following this approach, we use three ~8B models to rephrase MIMIC notes into "high-quality" text: Llama-3.1 (8B) (Dubey et al., 2024), Mistral-v0.3 (7B) (Jiang et al., 2023a), and Qwen-2.5 (7B) (Qwen et al., 2024). These efficient open-source LLMs also represent a realistic deployment choice for privacy- and resource-constrained clinical environments. Prompt templates, decoding parameters, and other inference details are provided in Appendix A. The complete process required about 1.8K GPU hours on NVIDIA H100 with vLLM (Kwon et al., 2023), producing a **9.1B-word** corpus, which we denote as MIMIC-EXT-SYNNOTES (details in Appendix Table 4).

4 Intrinsic Evaluation and Comparison with Human-written Notes

4.1 Experimental Setup

For intrinsic evaluation, we analyzed and compared surface-level and linguistic features between synthetic and human-written notes. We computed sentence and note lengths, part-of-speech (POS) distributions (for nouns, verbs, and numerals), and two standard readability scores: Flesch-Kincaid Grade Level (Solnyshkina et al., 2017) and SMOG Index (Harry and Laughlin, 1969).

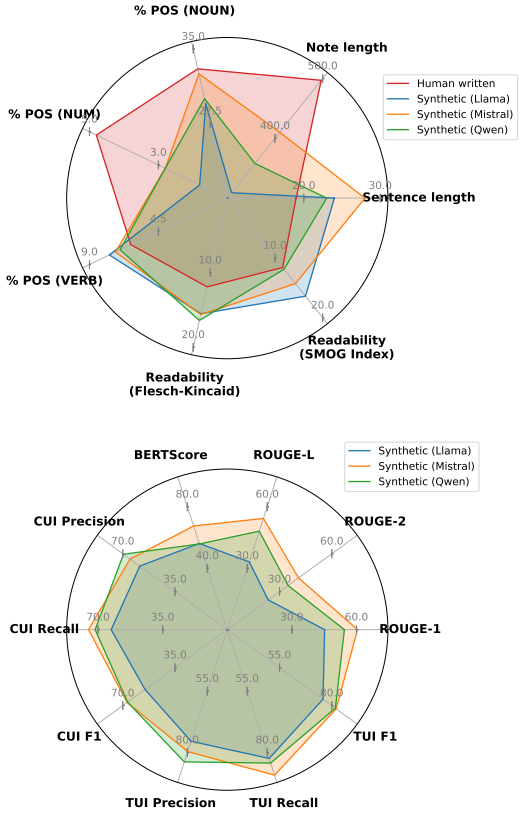


Figure 1: Linguistic features of synthetic notes and their textual similarities to human-written notes.

For pairwise metrics to capture the textual similarity between human-written and synthetic notes, we computed ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020). Due to token length limitations in many pre-trained BERTScore base models (typically capped at 512 tokens), we opted for a DeBERTa-based model (He et al., 2023) with flexible input length and adaptation to natural language inference, which offers enhanced capability for semantic similarity comparison.

To quantify the semantic fidelity of LLM-generated notes in terms of medical concepts, we followed previous works (Gao et al., 2022; Xu et al., 2024) to extract UMLS Concept Unique Identifiers (CUIs) using QuickUMLS (Soldaini, 2016), and computed precision, recall, and F1 scores using the original notes as the reference, as proposed in MedCON (Yim et al., 2023). CUIs capture unique medical concepts from UMLS as a bridge

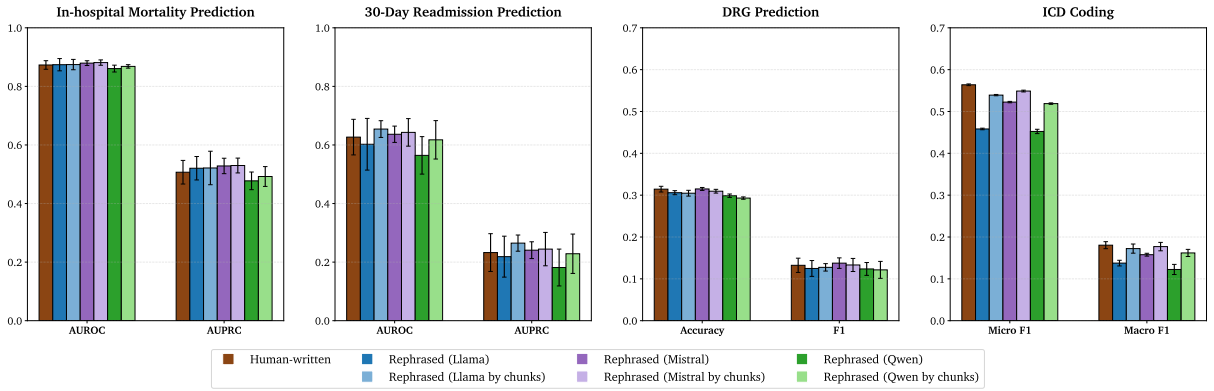


Figure 2: Comparing synthetic notes with human-written notes in downstream modeling. Each metric is calculated as the mean of five runs, and the error bars represent 95% confidence intervals.

for shared medical semantics. Following MedCON, we only considered clinically relevant semantic groups, such as anatomy, chemicals & drugs, and disorders. For a high-level comparison, we also mapped CUIs to their medical semantic types (Type Unique Identifiers, or TUIs) and computed the corresponding scores.

To facilitate efficient evaluation, we sampled 1K discharge summaries and 5K other note types from MIMIC-IV and MIMIC-III, respectively, as we observed that using more notes did not yield significantly different results.

4.2 Results

Figure 1 shows the comparison between human-written and LLM-generated notes. The synthetic notes are noticeably shorter in length but contain longer sentences with higher readability scores. They also show **a significant shift in lexical and semantic distributions**, as reflected by textual similarity scores with different degrees of variation across LLMs.

Analysis of TUI distributions (Table 1) reveals that the synthetic notes tend to omit concepts related to *"Body Space or Junction"* and *"Medical Device"* while introducing more from *"Natural Phenomenon or Process"* and *"Clinical Attribute"*, suggesting that synthetic text is **more descriptive but less dense with anatomical and specialist terminology**. This shows that LLMs may have a tendency to introduce more general terms and phrases in the generated text. The lists of frequent errors in semantic types made by each LLM are presented in Appendix Table 6.

	Semantic Type	%
Added	Natural Phenomenon or Process	16.5%
	Clinical Attribute	13.2%
	Phenomenon or Process	9.8%
	Pharmacologic Substance	9.5%
	Organ or Tissue Function	8.2%
Omitted	Body Space or Junction	12.2%
	Mental Process	11.8%
	Medical Device	11.6%
	Gene or Genome	11.2%
	Cell Component	8.8%

Table 1: Top 5 semantic types added (false positives) or omitted (false negatives) in synthetic notes, aggregated across three LLMs.

5 Extrinsic Evaluation with Healthcare Applications

5.1 Experimental Setup

We compared the utility of synthetic notes with human-written notes by training downstream clinical NLP models as a form of extrinsic evaluation. For a patient in a given task cohort, the input notes for the downstream model are either synthetic or human-written for both training and testing. In other words, we compared *"train synthetic, test synthetic" (TSTS)* and *"train real, test real" (TRTR)* by only varying the source of notes. The downstream performance is considered as a proxy to measure whether synthetic notes provide similar predictive utility to human-written notes. For this purpose, we did not consider *TSTR* (train synthetic, test real) or *TRTS* (train real, test synthetic), which can be confounded by factors such as writing styles and other features in the textual distribution.[1]

[1]Section 7.3 presents results for data augmentation, which aligns with *TSTR*.

We considered four clinical NLP tasks that represent key healthcare applications: risk stratification (in-hospital mortality and 30-day hospital readmission prediction) (Jiang et al., 2023b; Liang et al., 2025) and medical code assignment (DRG and ICD-10 prediction) (Wang et al., 2024a; Douglas et al., 2025). On the input side, they involve multiple note types authored by different caregivers at various stages of the care episode: in-hospital mortality (Jiang et al., 2023b; Si and Roberts, 2019) and DRG prediction (Wang et al., 2024a) tasks occur during the hospital episode and thus utilize all available clinical notes charted by the time of prediction (one day after ICU admission), such as nursing notes and radiology reports. The 30-day hospital readmission prediction (Boag et al., 2021; Huang et al., 2019) and ICD coding (Ji et al., 2024; Liu et al., 2022b) tasks occur after the end of hospital stay, and their input is based on the discharge summary alone.

On the output side, these tasks also involve varying complexities and requirements for clinical details: the first three tasks require the knowledge of overall patient status for a single-label classification, whereas the ICD coding task requires a higher level of attention to clinical details for the challenging multi-label classification with 8K ICD-10 codes (Edin et al., 2023). The four task cohorts are constructed from MIMIC with sample sizes ranging from 18K to 122K patients, and require processing of 546K notes. The statistics of the datasets are presented in Appendix Table 5.

For a given downstream task and a source of clinical notes, a classification model was trained, evaluated on the validation set for early stopping, and then tested on the held-out set to report final results. The only variation for a given task was the source of clinical notes, which are either human-written or synthetic, i.e., rephrased by one of the LLMs. Appendix B presents the details of the task cohorts and the training of downstream models.

5.2 Results

Figure 2 shows that synthetic notes exhibit comparable downstream performance to human-written notes on single-label tasks (i.e., all tasks except ICD coding), with results overlapping the 95% confidence intervals of using original notes. This suggests that **key clinical information is preserved in synthetic notes**, enabling accurate modeling and prediction for three prediction tasks. On the other hand, the ICD coding task (8K classes; multi-label) that requires fine-grained details exhibits much lower performance (with a drop of 4–10% in micro F1 score), suggesting **loss of certain clinical details** during synthesis that are essential for coding. Crucially, rephrasing *by chunks* instead of *by notes* substantially closes this performance gap, indicating **LLMs' difficulty in retaining and integrating information from long contexts** as a primary driver for detail loss.

5.3 Fairness Evaluation on Risk Stratification

Beyond predictive accuracy, we also assess fairness for two risk stratification tasks (in-hospital mortality and 30-day hospital readmission predictions). Following Brown et al. (2025), we split the patients into subgroups based on age (over and under 50), gender (female and male), and ethnicity (non-white and white), and computed the average absolute odds difference (AAOD) (Hardt et al., 2016) and statistical parity difference (SPD) (Besse et al., 2021) across these subgroups.

Figure 3 presents the results of the fairness evaluation, where we observe no significant changes in the fairness of the models trained on synthetic notes compared to those trained on human-written notes. This indicates that the LLMs did not introduce significant biases in their rephrasing that would affect downstream modeling of risk stratification.

6 Fact-checking with Error Analysis

We measure the truthfulness of the synthetic notes by fact-checking their claims against those made in the original notes. We then analyze the common error patterns exhibited by LLMs.

6.1 Experimental Setup

We follow Munnangi et al. (2025) to decompose synthetic notes into atomic facts (Min et al., 2023) and conduct fact-checking via entailment (Wang et al., 2024c). Following Xie et al. (2024), we calculate *precision* (the proportion of synthetic factual claims entailed by the human-written note) and *recall* (the proportion of reference factual claims entailed by those in the synthetic note).

For a pair of synthetic note N_S and human-written note N_H, they are decomposed into atomic factual claims C_S and C_H. Precision is defined as the proportion of facts in C_S that can be entailed by N_H:

$$Precision = \frac{1}{|C_S|} \sum_{c \in C_S} [N_H \vDash c] \quad (1)$$

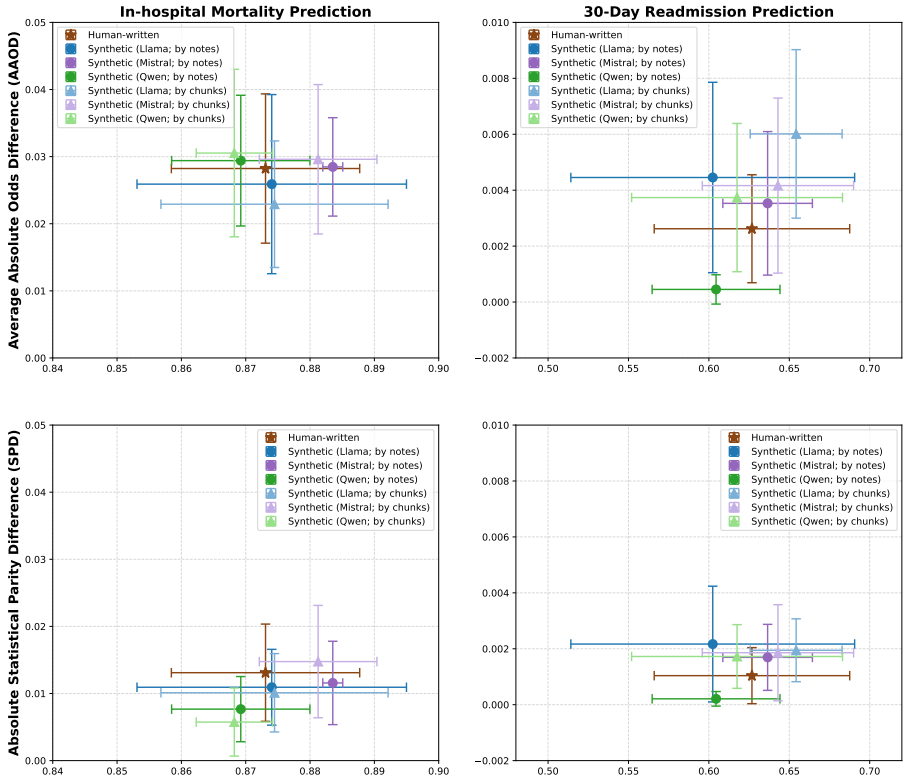


Figure 3: Fairness evaluation of the synthetic notes in downstream modeling in comparison with human-written notes. The x-axis presents model performance (AUROC score) and the y-axis presents model fairness (Average Absolute Odds Difference (AAOD) on the top panel and the absolute Statistical Parity Difference (SPD) on the bottom panel). For fairness metrics, the lower the better. The value of 0 for AAOD and SPD means no model discrimination between the examined subgroups. The results were averaged across age, gender, and ethnicity.

where $N_H \vDash c$ means note N_H entails fact c. Recall is defined as the proportion of facts in C_H that can be entailed by those from the synthetic notes C_S:

$$Recall = \frac{1}{|C_H|} \sum_{c \in C_H} [C_S \vDash c] \qquad (2)$$

Note that this is different from Munnangi et al. (2025), which used sentences from the reference note instead of decomposed facts; we found the latter to be more interpretable. The final precision and recall are macro averages over the sampled datasets, and F1 is the harmonic mean of the final precision and recall scores. We use Gemini-2.5 to decompose facts and assess entailment due to its strong reasoning capabilities. Due to the cost of querying the closed-source model to provide robust assessment, we conducted this part of the experiments on a small set of sampled notes as exploratory analysis, which covered more than 10K facts. More details are described in Appendix C.

6.2 Results

Table 2 shows that synthetic notes generally exhibit **high precision in their factual claims, but considerably lower recall**, indicating missing details. This becomes more pronounced when rephrasing *long notes* (>2K words), with recall as low as 0.3. Meanwhile, rephrasing *by chunk* produces a substantially larger number of facts, improving recall and suggesting a preservation of clinical details, consistent with findings in Section 5.

Beyond aggregated metrics, one author with a background in medical informatics analyzed sample outputs and identified a taxonomy of common error types (see Appendix Table 7): **misinterpretation of clinical context is the dominant failure mode**, followed by measurement errors, temporal/recency confusion, and fabrication of claims when the documentation is subtle. LLMs sometimes make up names and dates even though these are removed from real notes, and they also struggle with shorthand notation, e.g., confusing blood pressure values for respiration rates.

Finally, the results highlight a trade-off inher-

	Fact Prec	Fact Rec	Fact F1	# Avg. Facts
Llama	0.903	0.587	0.712	68
long notes	0.904	0.309	0.460	101
by chunk	0.868 (↓)	0.794 (↑)	0.830 (↑)	297 (↑)
Mistral	0.972	0.812	0.885	105
long notes	0.959	0.719	0.822	186
by chunk	0.938 (↓)	0.779 (↑)	0.851 (↑)	251 (↑)

Table 2: Fact-checking results based on decomposed atomic facts from synthetic notes.

ent to chunk-level rephrasing: while it maintains more claims and improves recall, it also lowers precision and is thus accompanied by higher hallucination rates. In many of these cases, LLMs see incomplete context and may generate plausible but unsupported content whose reference context appears in another chunk.

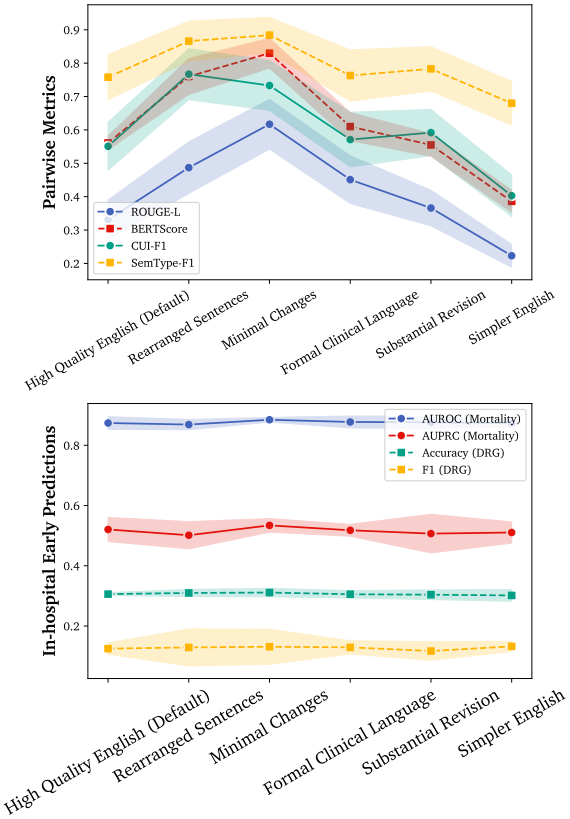


Figure 4: Impact of varied prompt instructions on intrinsic and extrinsic evaluations. Results are averaged over five random seeds.

7 Additional Analyses

We conduct additional analyses to examine the impact of prompting setups on the generated notes, including different instructions and iterative prompting. We also perform a data augmentation experiment on the ICD prediction task.

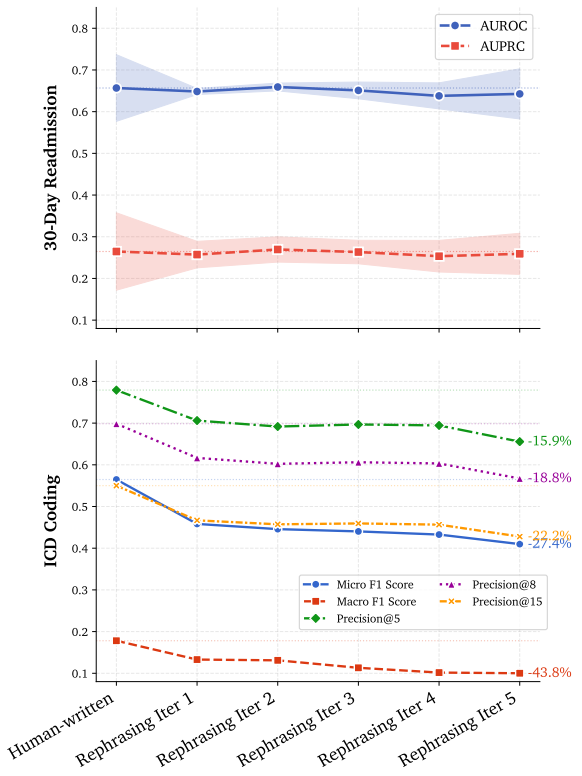


Figure 5: Impact of repeated synthesis on 30-day hospital readmission prediction across 5 iterations. Results are averaged over five random seeds.

7.1 Impact of Varied Prompts

We examine alternative prompts (Appendix E) that vary in structure (e.g., *"rearranging sentences"*) and style (e.g., *"simplified English"*), and observe their impact on synthetic notes, reflected by both intrinsic and extrinsic measurements. As expected, prompt variation significantly affects textual similarity: both syntactic and semantic contents change dramatically from instruction to instruction (Figure 4; top). Interestingly, the default instruction to produce *"high-quality"* language has a similar effect to *"substantial revision"*, showing the complexity of handling clinical documentation from the LLM's perspective.

On the other hand, the change of prompts has minimal impact on downstream modeling (Figure 4; bottom). This indicates that the synthetic notes preserve key clinical information related to patient severity despite differences in linguistic style, consistent with the findings in Section 5.

	All codes		Rare codes	
	micro F1	macro F1	micro F1	macro F1
Human-written notes	0.5644 (0.5622–0.5667)	0.1738 (0.1703–0.1772)	0.0359 (0.0173–0.0596)	0.0134 (0.0070–0.0199)
+*synthetic notes*	**0.5703** (0.5671–0.5736)	**0.2141** (0.2007–0.2278)	**0.1035** (0.0595–0.1485)	**0.0448** (0.0258–0.0656)

Table 3: Augmentation results for ICD coding. Models are evaluated on the human-written notes in the test set by bootstrapping 1000 samples from multiple seeds. The rare codes are 391 ICD codes that appeared only 10 times in the cohort (the benchmark in Edin et al. (2023) excluded codes that appeared fewer than 10 times). The reported scores are mean values and 95% confidence intervals. All augmented results show significant improvements with p-value <0.01 with two-sided t-test.

7.2 Impact of Repeated Synthesis

We explore repeated synthesis, where the LLM iteratively rephrases its own outputs to examine semantic drift. Shumailov et al. (2024) demonstrates cases where recursive generations can lead to model collapse when training happens in a loop. This raises critical concerns over the quality of data for training models over time, as generated content gradually becomes an integral part of the data pool, especially after deploying LLMs in practice.

Figure 5 shows the results on two downstream tasks over five iterations of repeated synthesis, each based on the output from the previous round. The performance on the single-label readmission prediction task remains stable. For ICD coding, although the performance decreases over time and sometimes by a large margin, the decline is progressive rather than sudden, and we have yet to observe a "collapse" in downstream modeling, i.e., scoring zero accuracy. Note that in this setting, the "generator" (i.e., the LLM) was kept unchanged, and only the downstream models (i.e., task-specific "predictor") were trained at each iteration instead of the "generator" (Shumailov et al., 2024). Moreover, the provision of a reference serves as a constraint to keep LLMs from generating completely irrelevant content. Nevertheless, the issue of LLMs missing fine-grained clinical details persists, as demonstrated by the declining ICD performance over repeated rephrasing.

7.3 Rephrased Notes for Augmentation

We examine the value of the synthetic notes in augmenting the training set for downstream modeling. Using LLMs to generate synthetic training data by following annotated examples has been studied for various clinical NLP tasks (Tang et al., 2023; Šuvalov et al., 2025). Unlike these studies, which craft application-specific prompts with demonstrations (i.e., few-shot prompting), our synthesis does not consider any task-related instructions or labels in the prompt. In other words, we apply the **task-agnostic synthetic notes for task-specific data augmentation**.

We focus on the challenging ICD-10 coding task and augment the original training set with synthetic notes generated by Mistral (by chunks), and evaluate on the human-written notes. Note that this *"test real"* setting is different from Section 5, where synthetic notes were examined in a *TSTS* setting.

The results are presented in Table 3, where we see that augmentation with synthetic notes increases all F1 scores (significant improvements with p-value <0.01), especially for rare codes (micro F1: $0.03 \rightarrow 0.10$). This shows the additional potential of LLM-synthesized clinical text for data augmentation despite being task-agnostic.

8 Discussion & Conclusion

Existing studies on synthetic clinical text typically evaluate the generation from one angle and lack direct comparisons between these evaluation strategies on a consistent set of clinical notes. We consider these different angles complementary in essence and useful for providing a unified view of the synthetic text quality; thus, we included textual similarity comparison, downstream utility assessment, and factuality checking in this study. As the results show, the synthetic clinical text generated by LLMs has both potential and weaknesses, and its usage should be assessed according to the application context.

Two findings from our study warrant particular attention for practitioners using LLMs for clinical text generation. **The first is the long-context bottleneck.** The substantial ICD coding gap between note-level and chunk-level rephrasing suggests the use of chunking to improve generation

quality. Meanwhile, the removal of surrounding context by chunking could lead to higher rates of hallucination. This means *a trade-off needs to be made on this decision based on the generative application and tolerance of each error type*, and future work needs to analyze in more depth the decisions related to chunking, such as boundary definition and chunk length.

Second, the error analysis shows that LLM failures are not random but exhibit recurring patterns. This suggests the potential for *future work to develop specialized techniques to address these errors*, such as better prompting for temporal anchoring and explicit handling of abbreviations.

We used rephrasing as the generation method in this study, and we did not explicitly distinguish between "rephrasing" and "synthesizing". Nevertheless, we acknowledge the nuanced difference between them in healthcare, especially the connotation of the term "synthesis" for privacy preservation. Grounding on real notes to create alternative versions of the clinical text — as done in "rephrasing" — forfeits privacy protection. Meanwhile, since the goal of the study is to measure text quality and to shed light on LLMs' capabilities, we maintain this decision to allow evaluation in different setups, i.e., head-to-head comparison with human-written notes and on downstream tasks.

Finally, our experiments show that the rephrased notes can be instrumental in understanding the quality of LLM generation and valuable for data augmentation despite being task-agnostic. To support future research, we compile and share the rephrased notes in this study as MIMIC-EXT-SYNNOTES, a synthetic corpus with 9B tokens based on 4.6M MIMIC notes.[2]

Limitations

Choice of LLMs for generation. We only focused on efficient open-source LLMs in our experiments. While more recent models may handle long contexts better, recent work suggests the gap may be narrower than expected: Kang et al. (2025) showed that using larger and more capable models for rephrasing pretraining data does not necessarily translate into better pretraining performance. Our complementary results with Gemma and Mixtral (Appendix F) did not yield substantially different results, supporting this view. Therefore, we consider the current model selection acceptable for fulfilling the study's objectives. Nevertheless, extending the analysis to models at other scales — both smaller and larger — and considering frontier models in a privacy-preserved setting are important future work.

Methods for text synthesis and evaluation. As mentioned, given the scope of this study, we did not explore alternative approaches for generating synthetic clinical text, particularly reference-free methods. Since MIMIC contains abundant EHR data, future work could incorporate structured measurements in a cross-modal setting to enhance the quality and validity of synthetic text. For evaluation methodologies, we consider the use of models in the current study — such as QuickUMLS for CUI extraction and BERT with sliding window for classification — acceptable as they are consistent for synthetic and human-written notes, but future work needs to further conduct meta-evaluation of the validity and reliability of these methods.

Issues with fact-checking. Although we conducted error analysis using LLM-as-judge for fact-checking, we were unable to directly compare synthetic notes with human-written notes. Such comparison requires extensive time and clinical expertise to review long notes and interpret patient care context, which was infeasible for this study. Instead, we opted to rely on downstream task performance and prior methods from fact-checking (Xie et al., 2024) that demonstrate strong correlation with expert judgment. We also acknowledge the limitations of our current fact-checking setup, which is based on a single family of models and a small number of notes, as we presented it as an exploratory analysis. The choice of model for LLM-as-judge particularly needs further scrutiny, given that models can be sensitive in reported factuality scores (Munnangi et al., 2025). Future work needs to validate the robustness of alternative models — probably the cheaper, open-source options — for this critical task. Nevertheless, we kept the use of models and notes consistent throughout the experiment, so that the analysis of errors and the comparison between note-level and chunk-level rephrasing are still valid.

Ethical Considerations

The generation of synthetic text using LLMs warrants careful ethical considerations. Although our synthesis was based on actual MIMIC notes and

[2]The synthetic note corpus can be found at https://github.com/JHLiu7/MIMIC-Ext-SynNotes.

offered generally high precision of factual claims, we observed that synthetic notes still omit or introduce information, which can be further exacerbated by insufficient context. This highlights the need for cautious use and, in some cases, additional filtering to ensure reliability, depending on the use case. Furthermore, while we assessed potential bias in downstream models for risk prediction, other dimensions of bias might remain and need to be considered in future work.

We assessed the utility of synthetic notes in downstream modeling by comparing *"train real, test real" (TRTR)* and *"train synthetic, test synthetic" (TSTS)*. Real-world signals (e.g., diagnosis codes) were utilized in both settings, so the *TSTS* models are applicable to *"test real"*. Meanwhile, careful consideration and transparency are essential if the synthetic notes were used to develop systems intended for implementation.

We followed the recommended practices for handling MIMIC notes. For rephrasing, we used only local, open-source models, and conducted all experiments in a secure environment to ensure data protection. For fact-checking with Gemini, we adhered to PhysioNet's Responsible Use guidelines[3] by accessing Gemini models via Vertex AI, ensuring compliance with HIPAA regulations.

Acknowledgments

We would like to thank our reviewers for their thoughtful, detailed, and constructive comments that helped to improve this paper. This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. This work was also supported by CSIRO (Australia's National Science Agency) through the Australian e-Health Research Centre.

References

Basel Alshaikhdeeb, Ahmed Abdelmonem Hemedan, Soumyabrata Ghosh, Irina Balaur, and Venkata Satagopam. 2025. Generation of synthetic clinical text: A systematic review. *arXiv [cs.CL]*.

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. 2021. A survey of bias in machine learning through the prism of statistical parity. *The American statistician*, pages 1–11.

William Boag, Olga Kovaleva, Thomas H McCoy, Jr, Anna Rumshisky, Peter Szolovits, and Roy H Perlis. 2021. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational psychiatry*, 11(1):32.

Katherine E Brown, Chao Yan, Zhuohang Li, Xinmeng Zhang, Benjamin X Collins, You Chen, Ellen Wright Clayton, Murat Kantarcioglu, Yevgeniy Vorobeychik, and Bradley A Malin. 2025. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association: JAMIA*, page ocaf038.

Kristine Ann M Carandang, Jasper Meynard Arana, Ethan Robert Casin, Christopher Monterola, Daniel Stanley Tan, Jesus Felix B Valenzuela, and Christian Alis. 2025. Are LLMs reliable? an exploration of the reliability of large language models in clinical note generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1413–1422, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A Lee, Caitlin E Coombes, Brad Bradshaw, Mahir A Sufian, Hyo Jung Hong, Teresa P Nguyen, Mohammad R Rasouli, Komal Kamra, and 10 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1).

Tiago K Colicchio, James J Cimino, and Guilherme Del Fiol. 2019. Unintended consequences of nationwide electronic health record adoption: Challenges and opportunities in the Post-Meaningful use era. *Journal of medical Internet research*, 21(6):e13313.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1):1–13.

James C Douglas, Yidong Gan, Ben Hachey, and Jonathan K Kummerfeld. 2025. Less is more: Explainable and efficient ICD code prediction with clinical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847, Stroudsburg, PA, USA. Association for Computational Linguistics.

[3]https://physionet.org/news/post/gpt-responsible-use

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2572–2582, New York, NY, USA. Association for Computing Machinery.

Yidong Gan, Maciej Rybinski, Ben Hachey, and Jonathan K Kummerfeld. 2025. Aligning AI research with the needs of clinical coding workflows: Eight recommendations based on US data analysis and critical review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 909–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M M Churpek, and Majid Afshar. 2022. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv [cs.CL]*.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

G Harry and Mc Laughlin. 1969. SMOG grading - a new readability formula. *The Journal of Reading*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv [cs.CL]*.

Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. 2024. A unified review of deep learning for automated medical coding. *ACM computing surveys*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7B. *arXiv [cs.CL]*.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, and 9 others. 2023b. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Lamin Juwara, Alaa El-Hussuna, and Khaled El Emam. 2024. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns (New York, N.Y.)*, 5(4):100946.

Nicolas M Kahl, Marshall J Frieden, Zach R Pope, Marlene M Millen, Vaishal M Tolia, Theodore C Chan, Christopher A Longhurst, Karandeep Singh, and Alan X You. 2026. Evaluation of electronic health record-integrated artificial intelligence chart review. *Npj Health Systems*, 3(1):6.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. 2025. Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10750–10769, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam, and Sven Gowal. 2024. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5148–5168.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, New York, NY, USA. ACM.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024. DataComp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rumeng Li, Xun Wang, and Hong Yu. 2023. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143, Singapore. Association for Computational Linguistics.

Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. 2025. Causal representation learning from multimodal clinical records under non-random modality missingness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28779–28796, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022a. "note bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of biomedical informatics*, 133:104149.

Jinghui Liu, Bevan Koopman, Nathan J Brown, Kevin Chu, and Anthony Nguyen. 2024a. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artificial intelligence in medicine*, 159(103027):103027.

Jinghui Liu and Anthony Nguyen. 2024. Rephrasing electronic health records for pretraining clinical language models. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 164–172.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2022b. Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of biomedical informatics*, 133:104161.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Mohammad Loni, Fatemeh Poursalim, Mehdi Asadi, and Arash Gharehbaghi. 2025. A review on generative AI models for synthetic medical text, time series, and longitudinal data. *npj digital medicine*, 8(1):281.

Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Charvi Bannur, Christina Baek, Darren Teh, David Schwab, Haakon Mongstad, Haoli Yin, Josh Wills, and 10 others. 2025. BeyondWeb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv [cs.LG]*.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-Tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore:

Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Stroudsburg, PA, USA. Association for Computational Linguistics.

Avijit Mitra, Zhichao Yang, Emily Druhl, Raelene Goodwin, and Hong Yu. 2025. Synth-SBDH: A synthetic dataset of social and behavioral determinants of health for clinical text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27887–27923, Stroudsburg, PA, USA. Association for Computational Linguistics.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle A Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi Omiye, Mehr Kashyap, and Nigam Shah. 2025. FactEHR: A dataset for evaluating factuality in clinical notes using LLMs. In *Machine Learning for Healthcare Conference*. PMLR.

Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. 2024. e-health CSIRO at RRG24: Entropy-augmented self-critical sequence training for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 99–104.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv [cs.CL]*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *arXiv [cs.CL]*.

Kirk Roberts. 2024. Large language models for reducing clinicians' documentation burden. *Nature medicine*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):1–39.

Arne Schwieger, Katrin Angst, Mateo de Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, David Troi, Erich Seifritz, Samantha Weber, and Sebastian Olbrich. 2024. Large language models can support generation of standardized discharge summaries - a retrospective study utilizing ChatGPT-4 and electronic health records. *International journal of medical informatics*, 192(105654):105654.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Yuqi Si and Kirk Roberts. 2019. Deep patient representation of clinical notes via Multi-Task learning for mortality prediction. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:779–788.

William R Small, Jonathan Austrian, Luke O'Donnell, Jesse Burk-Rafel, Katherine A Hochman, Adam Goodman, Jonah Zaretsky, Jacob Martin, Stephen Johnson, Vincent J Major, Simon Jones, Christian Henke, Benjamin Verplanke, Jwan Osso, Ian Larson, Archana Saxena, Aron Mednick, Choumika Simonis, Joseph Han, and 12 others. 2025. Evaluating hospital course summarization by an electronic health record-based large language model. *JAMA network open*, 8(8):e2526339.

Luca Soldaini. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. *MedIR workshop, sigir*.

M Solnyshkina, R Zamaletdinov, L Gorodetskaya, and A Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of social studies education research*, 8(3):238–248.

Sarvesh Soni and Dina Demner-Fushman. 2024. Toward relieving clinician burden by automatically generating progress notes using interim hospital data. *AMIA Annual Symposium Proceedings*, 2024:1059–1068.

Sarvesh Soni and Kirk Roberts. 2020. Paraphrasing to improve the performance of electronic health records question answering. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2020:626–635.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of LLMs help clinical text mining? *arXiv [cs.CL]*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2024a. DRG-LLaMA : tuning

LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ digital medicine*, 7(1):16.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024b. NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024c. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. DocLens: Multi-aspect fine-grained medical text evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Stroudsburg, PA, USA. Association for Computational Linguistics.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

Wen-Wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, Curtis P Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 0(0).

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using synthetic health care data to leverage large language models for named entity recognition: Development and validation study. *Journal of medical internet research*, 27(1):e66279.

A Details of Rephrasing MIMIC Notes

For clinical text synthesis, each LLM was provided with a system prompt consistent with its chat template and was instructed to synthesize a new version using a paraphrasing prompt found effective in prior studies (Liu and Nguyen, 2024; Maini et al., 2024), denoted as the "high quality" prompt:

Rephrasing prompt

```
[System Prompt] You are a medical artificial
 intelligence assistant. The assistant gives
 truthful, detailed, and professional
answers to the requests.

For the following paragraph give me a
diverse paraphrase of the same in high-
quality English language as in clinical
notes written by medical professionals:
```

We consider two rephrasing setups, one *by note* and the other *by chunk*. For rephrasing *by note*, we stratify the clinical notes by lengths — $(0, 500]$, $(500, 1000]$, $(1000, 2000]$, $(2000, 4000]$, and > 4000 — and set the corresponding maximum number of generated tokens to 1000, 2000, 4000, 8000, and 10,000, respectively. For rephrasing *by chunk*, we set the maximum number of generated tokens to 512 given the short chunk length (~150 words). Temperature was set to 0.75 and top-p 0.9 for all LLMs.

B Implementation Details on Extrinsic Evaluation

We followed previous studies (Liu et al., 2022a) to construct the in-hospital prediction cohorts (mortality and DRG) from MIMIC-III. For ICD coding, we used the benchmark (Edin et al., 2023) that comprehensively examined various types of ICD codes from MIMIC and focused on ICD-10 based on MIMIC-IV. Only ICD codes with a frequency higher than 10 were retained in the benchmark. This threshold was also applied to the DRG cohort with MS-DRG (Medicare Severity-DRG). We constructed a readmission cohort from MIMIC-IV by considering only the first admission of each patient and discarding immediate readmission cases that occurred within 24 hours after discharge. The statistics of the datasets are presented in Table 5.

We chose PLM-ICD (Huang et al., 2022; Edin et al., 2023) as the classification model given its strong performance on ICD coding and adaptability to handle long texts with sliding window. More

	# Total words (billion)	# Avg words per note	# Avg words per chunk
Llama (by notes)	1.21	260	-
Mistral (by notes)	1.34	289	-
Qwen (by notes)	1.13	263	-
Llama (by chunks)	2.24	482	206
Mistral (by chunks)	1.74	376	160
Qwen (by chunks)	1.41	315	146
Total	9.1	332	-

Table 4: Statistics of MIMIC-EXT-SYNNOTES generated by LLMs. When rephrasing based on chunks, each note from MIMIC is segmented into 2.3 chunks on average. The synthetic notes contain 9.1B words in total.

specifically: 1) it leverages a pre-trained encoder (BERT), allowing for domain-specific initialization; 2) it processes text via a sliding-window mechanism, making it efficient for long clinical documents; 3) it has demonstrated strong performance, particularly for ICD coding, compared to other deep learning methods (Edin et al., 2023); and 4) it is adaptable across tasks by modifying only the classification layer.

We initialized the model with a biomedical RoBERTa checkpoint (Lewis et al., 2020). Training was conducted using a batch size of 16 for up to 10 epochs, with early stopping triggered after no improvement for 3 epochs. The learning rate was set to 5e-5, with 2,000 warmup steps followed by linear decay. This training recipe was applied to all versions of input notes, either human-written or LLM-rephrased, to ensure consistent comparisons. The trained model was then evaluated on the test set of the corresponding task. We averaged the results over five runs with different random seeds to report the main results.

C Fact-checking Implementation

Following previous works (Munnangi et al., 2025; Xie et al., 2024), we performed fact-checking of the synthetic notes by decomposing the synthetic notes into atomic facts and measuring entailment. We use Gemini-2.5-Pro, a thinking model, to decompose a note into atomic facts, and Gemini-2.5-Flash to assess entailment. We used the prompts from Munnangi et al. (2025) for fact decomposition and entailment judgment, which are presented in the following.

We sampled 25 notes of varied length (10 radiology reports and 15 discharge summaries) for fact-checking due to the cost related to querying Gemini. We also consider *long notes* over 2K words, constituting five discharge summaries from the 25 notes. We show results (Table 2) on both these 25 notes and the five long notes. For synthetic notes rephrased *by chunk*, we also only present the results on the five long notes. Due to the prolonged content, these five notes take significantly longer to decompose and judge, sometimes requiring 500+ API calls for a single pair. Given the cost consideration, we relied on this small subset for this section. We also only focused on Llama and Mistral for this analysis, which contributed 5.2K and 6.1K facts, respectively.

D Omitted and Added Semantic Types

Table 6 presents the top-10 semantic types of the medical concepts that are either omitted or added by the LLMs after rephrasing. These are based on MedCON (Yim et al., 2023) but mapped CUI to TUI. The added semantic types can be considered as false positives (FPs) and omitted ones as false negatives (FNs) when comparing against the human-written notes.

E Alternative Prompts

We consider the following alternative prompts to explore their impact on generated text.

Rearranged Structure *Rewrite the following paragraph by rearranging sentences and phrases, while retaining the original clinical content and meaning:*

Minimal Changes *Rephrase the following paragraph with minimal changes, keeping the original structure and meaning intact. Make only slight adjustments for clarity and conciseness while preserving the clinical details:*

Formal Clinical Language *Paraphrase the following paragraph using strictly formal clinical documentation language, maintaining the same level of detail and medical accuracy:*

	# Patients	% Train/Val/Test	# Classes	Input notes	Metrics
In-hospital Mortality	33,515 (MIMIC-III)	80%/10%/10%	2 (11.1%)	Nursing notes, physician notes, etc.	AUROC AUPRC
30-Day Readmission	33,518 (MIMIC-IV)	80%/10%/10%	2 (13.0%)	Discharge summary	AUROC AUPRC
DRG Prediction	18,681 (MIMIC-III)	70%/20%/10%	369	Nursing notes, physician notes, etc.	Accuracy F1
ICD Coding	122,278 (MIMIC-IV)	73%/16%/11%	7,942	Discharge summary	Micro F1 Macro F1

Table 5: Description of cohorts and prediction tasks. For diagnosis-related classifications, DRG prediction is based on MS-DRG and ICD coding on ICD-10. All tasks have a single label per patient except for ICD coding, which is multi-label multi-class classification.

Substantial Revision *Produce a substantially revised version of the following paragraph in professional clinical English, changing sentence structure and vocabulary while preserving all critical medical details:*

Simpler English *Paraphrase the following paragraph into simpler clinical English, minimizing jargon and long sentences but keeping the key medical details:*

Entailment judgment prompt

```
You are an expert on natural language
entailment.
Your task is to deduce whether premise
statements entail hypotheses.
Return only '1' if the hypothesis can be
fully entailed by the premise.
Return only '0' if the hypothesis contains
information that cannot be entailed by the
premise.
Generate the answer in JSON format with the
following keys:
'entailment_prediction': 1 or 0, whether the
 claim can be entailed.
Only return the JSON-formatted answer and
nothing else.

Premise: {{premise}}
Hypothesis: {{hypothesis}}

Here is the JSON-formatted answer:
```

Fact decomposition prompt

```
Please breakdown the following text into
independent facts as a string delimited by
"//" to separate the facts

Example 1:
Note: "There is a dense consolidation in the
 left lower lobe."

Atomic facts:
There is a consolidation. // The
consolidation is dense. // The consolidation
 is on the left. // The consolidation is in
a lobe. // The consolidation is in the lower
 portion of the left lobe.

Example 2:

Note: "The patient has been having
intermittent shortness of breath for the
last two years."

Atomic facts:
The patient has been having shortness of
breath. // The shortness of breath is
intermittent. // The shortness of breath has
 been present for the last two years.

Do not include any other text, or say "Here
is the list..."

Note: {{note}}
```

F Alternative Models

Table 8 presents the numerical results presented in Figure 2. We provide additional results on rephrasing with Gemma-2 (9B) and Mixtral-v0.1 (8×7B) by chunks with one seed.

Synthetic text rephrased by Llama	
Added Semantic Types (FP) *Total counts: 9,321*	*Omitted Semantic Types (FN)* *Total counts: 8,979*
Natural Phenomenon or Process (19.1%)	Medical Device (13.0%)
Clinical Attribute (15.2%)	Body Space or Junction (12.4%)
Organ or Tissue Function (10.9%)	Mental Process (11.7%)
Pharmacologic Substance (9.9%)	Biomedical or Dental Material (11.3%)
Phenomenon or Process (8.1%)	Gene or Genome (9.9%)
Organism Attribute (7.9%)	Body System (9.1%)
Body Substance (7.6%)	Indicator, Reagent, or Diagnostic Aid (8.8%)
Organism Function (7.4%)	Cell Component (8.5%)
Laboratory or Test Result (7.3%)	Organ or Tissue Function (7.9%)
Body Part, Organ, or Organ Component (6.6%)	Physiologic Function (7.3%)
Synthetic text rephrased by Mistral	
Added Semantic Types (FP) *Total counts: 7,409*	*Omitted Semantic Types (FN)* *Total counts: 4,238*
Natural Phenomenon or Process (15.9%)	Gene or Genome (14.6%)
Clinical Attribute (11.9%)	Body Space or Junction (13.0%)
Phenomenon or Process (11.4%)	Mental Process (11.3%)
Organ or Tissue Function (11.2%)	Medical Device (10.3%)
Organism Attribute (9.9%)	Cell Component (9.4%)
Organism Function (8.5%)	Biomedical or Dental Material (8.9%)
Body Space or Junction (8.3%)	Biologic Function (8.8%)
Laboratory or Test Result (8.2%)	Body System (8.0%)
Pharmacologic Substance (7.8%)	Indicator, Reagent, or Diagnostic Aid (7.9%)
Body Substance (6.9%)	Organ or Tissue Function (7.8%)
Synthetic text rephrased by Qwen	
Added Semantic Types (FP) *Total counts: 5,641*	*Omitted Semantic Types (FN)* *Total counts: 5,288*
Natural Phenomenon or Process (13.1%)	Mental Process (12.4%)
Clinical Attribute (11.9%)	Body Space or Junction (11.4%)
Pharmacologic Substance (11.0%)	Clinical Attribute (11.2%)
Phenomenon or Process (10.4%)	Gene or Genome (10.7%)
Cell (9.8%)	Medical Device (10.3%)
Element, Ion, or Isotope (9.5%)	Body Location or Region (9.3%)
Laboratory or Test Result (8.8%)	Nucleic Acid, Nucleoside, or Nucleotide (9.3%)
Molecular Function (8.6%)	Cell Component (8.6%)
Organism Function (8.6%)	Indicator, Reagent, or Diagnostic Aid (8.4%)
Biomedical or Dental Material (8.3%)	Organism Attribute (8.4%)

Table 6: Changes of semantic types in the rephrased notes by the three LLMs. Top 10 semantic types are listed for false positive (FP) and false negative (FN) compared to the original, human-written notes. The ratio is calculated based on total counts of the presented semantic types for each LLM for FP and FN, respectively.

Claim	Error Category	Comment
"The patient presented with fever."	No error	*This is a correct claim.*
"The patient continued to have significant oxygen requirements."	No error	*Correct.*
"An MRI later confirmed the focus was a subacute infarction."	Misinterpretation	*This is likely, but not confirmed.*
"The anemia was treated with iron supplementation."	Fabricated claim	*Patient had iron-binding study. No supplementation intake was documented.*
"The surgery occurred on discharge day."	Misinterpretation	*Wrong interpretation of "Surgery Discharge part".*
"The patient was cleared by occupational therapy (OT) for discharge to rehabilitation."	Misinterpretation	*Cleared by PT.*
"The patient was scheduled for a follow-up appointment with her gastroenterologist."	Misinterpretation	*Appointment scheduled with primary care physician.*
"The patient's name is Mr. John Doe."	Fabricated claim	*Made-up name.*
"The patient is a non-smoker."	Misinterpretation	*Incorrect; tobacco use is mentioned.*
"The patient consumes approximately 1–2 drinks per day."	Measurement/Numeracy	*EtOH: [**4–20**] drinks/day.*
"The adjusted medications include simvastatin."	Fabricated claim	*No adjustment.*
"The patient is asymptomatic with the low blood sugar levels."	Misinterpretation	*"He states that he is asymptomatic," but note shows symptomatic hypoglycemia.*
"In the emergency department, the patient's vital signs were stable."	Misinterpretation	*Incorrect; labs showed elevated WBC.*
"The patient's most recent pulse was 95 bpm."	Temporal/Recency	*Not most recent.*
"The patient's most recent urine output was 2 L."	Temporal/Recency	*Not most recent.*
"The patient is a former smoker."	Temporal/Recency	*Current smoker.*
"All pathology samples of bone and cartilage showed hyaline cartilage with focal acute inflammation."	Misinterpretation	*Incorrect summarization of pathology results: not all samples show hyaline cartilage with focal acute inflammation.*
"In the ER, the patient's respiration rate was 142/min."	Measurement/Numeracy	*142 is blood pressure.*
"In the ER, the patient's blood pressure was 18/97 mmHg."	Measurement/Numeracy	*18 is respiratory rate; 96 (instead of 97) is oxygen saturation.*
"The patient's discharge blood pressure was 168/45 mmHg."	Misinterpretation	*BP in ED, not discharge.*

Table 7: Categorization and analysis of example false positive claims in synthetic texts.

	In-hospital Mortality		30-day Readmission		ICD Coding		DRG Classification	
Input source	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**Micro F1**	**Macro F1**	**Accuracy**	**F1**
Human-written	0.873 (0.858–0.888)	0.507 (0.467–0.547)	0.627 (0.566–0.688)	0.233 (0.168–0.297)	0.564 (0.562–0.566)	0.180 (0.172–0.189)	0.314 (0.308–0.321)	0.132 (0.115–0.149)
Synthetic text *by notes*								
Llama	0.874 (0.853–0.895)	0.521 (0.480–0.561)	0.602 (0.514–0.691)	0.219 (0.149–0.289)	0.458 (0.456–0.460)	0.138 (0.131–0.144)	0.306 (0.301–0.311)	0.125 (0.106–0.144)
Mistral	0.879 (0.871–0.888)	0.528 (0.502–0.555)	0.637 (0.609–0.664)	0.241 (0.212–0.269)	0.522 (0.521–0.524)	0.158 (0.154–0.161)	0.315 (0.311–0.318)	0.138 (0.125–0.150)
Qwen	0.861 (0.849–0.872)	0.477 (0.447–0.508)	0.564 (0.500–0.628)	0.182 (0.119–0.244)	0.452 (0.447–0.458)	0.122 (0.110–0.135)	0.299 (0.295–0.303)	0.124 (0.108–0.139)
Synthetic text *by chunks*								
Llama	0.874 (0.857–0.892)	0.521 (0.464–0.579)	0.654 (0.626–0.683)	0.265 (0.237–0.293)	0.539 (0.538–0.541)	0.172 (0.161–0.183)	0.305 (0.298–0.312)	0.127 (0.119–0.136)
Mistral	0.881 (0.872–0.890)	0.530 (0.504–0.555)	0.643 (0.596–0.690)	0.244 (0.187–0.301)	0.549 (0.547–0.551)	0.177 (0.167–0.187)	0.309 (0.305–0.314)	0.133 (0.118–0.149)
Qwen	0.868 (0.862–0.874)	0.492 (0.458–0.526)	0.618 (0.552–0.683)	0.229 (0.161–0.296)	0.519 (0.517–0.521)	0.162 (0.153–0.170)	0.293 (0.290–0.296)	0.121 (0.101–0.142)
Additional synthesis *by chunks*								
Gemma	0.870	0.512	0.654	0.285	0.537	0.171	0.293	0.109
Mixtral	0.882	0.548	0.662	0.275	0.548	0.169	0.314	0.123

Table 8: Downstream modeling results for the four tasks based on different sources of clinical notes. Values are shown as mean with 95% confidence intervals in parentheses. The results for the three main LLMs are also presented in Figure 2.

Bridging the Version Gap: Multi-version Training Improves ICD Code Prediction, Especially for Rare Codes

Jinghui Liu **Anthony Nguyen**
Australian e-Health Research Centre, CSIRO
{jinghui.liu,anthony.nguyen}@csiro.au

Abstract

Clinical coding maps clinical documentation to standardized medical codes, an essential yet time-consuming administrative task that could benefit from automation. Current models on ICD coding are typically optimized for codes from a specific ICD version. However, in reality, ICD systems evolve continuously, and different versions are adopted across time periods and regions. Moreover, ICD coding suffers from the long-tail problem, and rare code performance can be a bottleneck for developing implementable models. We examine whether it is viable to train version-independent models by combining data annotated in different ICD versions, which may help address these challenges. We add ICD-9 data to the training of a modified label-wise attention model for ICD-10 prediction, and find that despite the version mismatch, adding ICD-9 yields a 27% increase in micro F1 for 18K rare ICD codes compared to training on ICD-10 alone. On 8K frequent ICD-10 codes, the multi-version training also substantially improves macro metrics, with far fewer model parameters.

1 Introduction

The International Classification of Diseases (ICD) is the global standard for reporting health conditions and diseases. It plays a critical role in healthcare billing, epidemiological research, and health policymaking (Dong et al., 2022; Gan et al., 2025). Assigning ICD codes to clinical documents requires a high level of expertise given that the candidate code set is often large (e.g., ICD-10 includes over 70K diagnosis codes) and the documentation can be lengthy and complex (Motzfeldt et al., 2025; Liu et al., 2023). Consequently, the coding process is time-consuming even for experienced coders, which has motivated extensive research into NLP models that automate or assist the coding process (Stanfill et al., 2010; Ji et al., 2024).

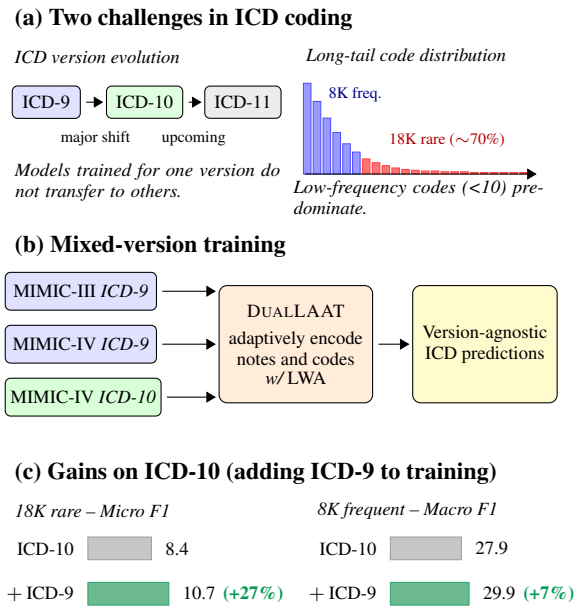


Figure 1: **(a)** ICD coding faces two intertwined challenges: the ICD system evolves continuously and the code distribution is heavily long-tailed. **(b)** We mix three MIMIC-derived datasets spanning ICD-9 and ICD-10 to train a single version-agnostic model, DUALLAAT. **(c)** Adding ICD-9 to ICD-10 training yields a 27% relative gain in micro F1 on rare ICD-10 codes than training on ICD-10 alone.

Current best-performing ICD coding models typically frame the task as a multi-label, multi-class classification problem (Douglas et al., 2025; Edin et al., 2023; Yuan et al., 2022), which requires defining a fixed label space and a substantial amount of data for supervised training. This setup faces two critical challenges in real-world ICD coding. First, **the ICD coding system undergoes continuous updates**, with new codes added and old codes retired. For example, CMS actively updates ICD-10 every year to keep codes aligned with healthcare needs.[1] At certain points, the ICD system undergoes major

[1] https://www.cms.gov/medicare/coding-billing/icd-10-codes

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 372–381
July 3-4, 2026 ©2026 Association for Computational Linguistics

transitions, such as the shift from ICD-9 to ICD-10, and the anticipated move to ICD-11 (Harrison et al., 2021). Models with a pre-defined label space have trouble handling these updates, which introduce substantial data shifts (Finlayson et al., 2021).

Second, **ICD code distributions are long-tailed** (Li et al., 2025a; He et al., 2025), which is a critical modeling challenge as the ICD system grows larger and more fine-grained. For example, in MIMIC-IV (Johnson et al., 2023), infrequent ICD-10 codes that appear fewer than 10 times (18K) account for 69.6% of all codes in the database (26K). Collecting sufficient samples for these rare and underrepresented diseases is essential for training. However, achieving this requires consolidating patient records across time periods and regions, which is likely accompanied by mismatched ICD versions. It remains understudied whether such a combination leads to benefits or harm in modeling. For example, given that only 24.3% of ICD-9-CM codes have exact matches in ICD-10-CM (Fung et al., 2021), mixing multiple datasets across versions may plausibly introduce noise that harms rather than helps modeling.

This study explores this question by examining if ICD-9 data is valuable for training an ICD-10 prediction model (Figure 1). Our experiments used two ICD-9 datasets and one ICD-10 dataset based on MIMIC (Edin et al., 2023), and mixed their training sets to train a version-independent model based on label-wise attention (LWA) (Wu et al., 2024; Mullenbach et al., 2018). Adding ICD-9 data to the training set led to **substantially improved performance on rare codes** (micro F1: $8.4 \rightarrow 10.7$) in ICD-10 using **a much smaller model**, as well as **improved per-label performance on frequent codes** (macro F1 of 29.9). Our findings demonstrate the synergistic effect of combining multi-version datasets for ICD coding, showcasing the potential value of leveraging legacy patient records to train more robust ICD prediction systems.

2 Related Work

Research on automated and assisted ICD coding dates back to the 1990s (Larkey and Croft, 1996). Recent advances have largely focused on configuring neural architectures to better model the large code set (Li et al., 2025b; Shi et al., 2017). Ji et al. (2024) surveyed modeling approaches and highlighted that label-wise attention (LWA) (Mullenbach et al., 2018; Wu et al., 2024; Vu et al., 2020) remains a key module to train high-performing models. Edin et al. (2023) conducted a comprehensive benchmark of various LWA-based models across three datasets spanning MIMIC-III and IV and ICD-9 and ICD-10, finding that PLM-ICD with in-domain pretrained checkpoints performed best. Other approaches exploit the relations between codes in the ICD ontology to improve modeling (Luo et al., 2024; Yuan et al., 2022). To the best of our knowledge, all prior models were trained for a specific set of ICD codes as label space, and none were evaluated across different ICD versions.

A growing number of studies have explored prompting LLMs to predict ICD codes (Simmons et al., 2025; Yang et al., 2023). While simple prompting achieves low performance (Soroush Ali et al., 2024), agent-based systems for ICD prediction show promise (Motzfeldt et al., 2025; Zheng et al., 2025). The generative approach is not constrained by a fixed label space and may offer enhanced interpretability. However, most of these studies were conducted on small-scale datasets with a limited label space of about 1K (Cheng et al., 2023), which contrasts with existing benchmarks (8K in Edin et al. (2023)) and public databases (26K in Johnson et al. (2023)). In addition, agent-based approaches often require large token budgets, which can be costly. For this study, we focus on comparing with supervised approaches on datasets with large code sets.

3 Methods and Experiments

3.1 Benchmark Datasets

We adopted the benchmark from Edin et al. (2023), which includes three MIMIC-derived datasets, namely MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10). The original benchmark focuses on frequent codes, defined as appearing $\geq$ 10 times in the corresponding database. To also evaluate rare codes, we re-extracted all codes from MIMIC, resulting in substantial expansions of the code space ($6K \rightarrow 11K$ for ICD-9 and $8K \rightarrow 26K$ for ICD-10). Since ICD-9 is no longer in active use, we focused on ICD-10 as the primary evaluation benchmark. Dataset statistics are presented in Appendix Table 5.

3.2 Classification Model

LWA trains code-specific embeddings to attend to relevant text tokens in clinical notes and uses them for classification (Edin et al., 2024). To enable

adaptive modeling across ICD versions, we trained an encoder to generate code embeddings from code descriptions, which are unique textual strings describing each ICD code. We jointly trained two text encoders – one for clinical notes and one for ICD codes – using LWA. Given the dual-encoder feature and the use of LWA from LAAT (Vu et al., 2020), we denote the model as DUALLAAT. The model takes N notes and C codes (as text strings) as input and outputs a probability matrix $\hat{\mathbf{y}} \in \mathbb{R}^{N \times C}$ as prediction. The codes in C are version-agnostic and can be ICD-9 or ICD-10. Detailed notations and training procedures are provided in Appendix B.

3.3 Experiments

To investigate the impact of mixing multiple ICD data sources, we aggregated the three train sets from MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10) for training and evaluated on the corresponding test sets. We included all ICD codes (i.e., from full label space) in the aggregated data to train the model. To enable fast experimentation and to highlight the impact of data mixing, we used basic CNN (Mullenbach et al., 2018) and RNN (Vu et al., 2020) as text encoders, denoted as $\text{DUALLAAT}_{\text{cnn}}$ and DUALLAAT, respectively.

We adopted the benchmark metrics (Edin et al., 2023) for evaluation on frequent codes, including both classification and ranking metrics. In particular, macro F1 was calculated as the arithmetic mean of F1 scores per code. Higher macro F1 indicates stronger performance on less frequent codes, reflecting tail-end robustness. We also evaluated on full ICD-10 set and explicitly on rare codes, which in practice can be equally important as frequent codes for reporting disease prevalence and enabling appropriate reimbursement.

3.4 Baselines

We consider two strong supervised models as baselines for frequent and full code sets, respectively:

PLM-ICD. This model adopts in-domain BERT with sliding-window as the text encoder and trains with LWA (Huang et al., 2022). Prior work found replacing BERT with Llama on a dataset with 50 codes did not improve performance (Motzfeldt et al., 2025), highlighting the strong baseline performance of PLM-ICD. We report its results from the previous benchmark study on the frequent 8K codes (Edin et al., 2023).

CoRelation. This model improves code representation learning by modeling the relations between codes through a bipartite graph for enhanced accuracy and efficiency (Luo et al., 2024). It was evaluated on top-50 and full codes in MIMIC. We use this model for reference comparison on full ICD-10 set with 26K codes.

4 Results

4.1 Impact of Data Mixing on Frequent Code Prediction and Benchmark Comparison

Table 1 presents the results on frequent codes. On ICD-10 prediction, mixing additional ICD-9 datasets from MIMIC improves all metrics for both $\text{DUALLAAT}_{\text{cnn}}$ and $\text{DUALLAAT}_{\text{rnn}}$, with about 2 points of improvement on macro F1. This shows that adding ICD-9 to training is beneficial for ICD-10 prediction despite terminological differences. Using a stronger encoder (RNN over CNN) also brings steady improvements, consistent with prior findings (Edin et al., 2023) and suggesting potential gains with stronger encoder models.

Compared with the PLM-ICD baseline, RNN-based DUALLAAT trained on mixed MIMIC datasets achieved similar precision but much higher macro AUC-ROC and macro F1. We also report results for the weaker CAML (Mullenbach et al., 2018) and LAAT (Vu et al., 2020) in Appendix Table 6. These baseline methods with CNN and RNN encoders form a more direct comparison with our model. The additional results also show that DUALLAAT trained using mixed datasets offers lower variance across the metrics.

4.2 Impact on Rare Code Prediction

Mixing additional data sources shows substantial benefits in rare code prediction, as presented in Table 2. Both ICD-10 and 9 benefited from alternative ICD data. For ICD-10, this yielded a 27.4% increase in micro F1, which is a substantial improvement given the number of codes (18K).

Table 4 compares the model performance on the full code set with 26K labels from MIMIC-IV ICD-10. DUALLAAT achieves results comparable to the reported CoRelation numbers on shared metrics, with a notably higher macro F1 (6.3→11.2), suggesting stronger performance across the long tail of rare codes. This shows that enhancing training data can be as important in ICD coding as architectural improvements, which have been the predominant focus in ICD prediction literature (Ji et al., 2024; Dong et al., 2022).

	Classification				Ranking			
	AUC-ROC		F1		Precision@k		R-precision	MAP
	Micro	Macro	Micro	Macro	8	15		
MIMIC-III ICD-9								
PLM-ICD (Edin et al., 2023)	98.9	95.9	59.6	26.6	72.1	56.5	60.1	64.6
DUALLAAT	**99.2**	**96.9**	**61.8**	**33.3**	**74.9**	**58.8**	**62.8**	**68.1**
MIMIC-IV ICD-9								
PLM-ICD (Edin et al., 2023)	99.4	97.2	62.6	29.8	70.0	53.5	62.7	68.0
DUALLAAT	**99.5**	**97.6**	**63.4**	**34.3**	**71.0**	**54.3**	**63.8**	**69.3**
MIMIC-IV ICD-10								
PLM-ICD (Edin et al., 2023)	99.2	96.6	**58.5**	21.1	**69.9**	**55.0**	57.9	61.9
DUALLAAT$_{cnn}$ (ICD-10 only)	99.2	96.8	55.6	24.3	67.5	52.7	55.3	59.0
DUALLAAT$_{cnn}$	99.3 (↑)	97.2 (↑)	56.6 (↑)	26.0 (↑)	68.4 (↑)	53.6 (↑)	56.4 (↑)	60.4 (↑)
DUALLAAT (ICD-10 only)	99.3	97.1	57.5	27.9	69.3	54.4	57.2	61.5
DUALLAAT	**99.3** (–)	**97.4** (↑)	58.0 (↑)	**29.9** (↑)	**69.9** (↑)	54.9 (↑)	**58.0** (↑)	**62.3** (↑)

Table 1: Results on frequent ICD codes for MIMIC-III *ICD-9* (3.7K), MIMIC-IV *ICD-9* (6K) and MIMIC-IV *ICD-10* (8K). The means of three random seed runs were reported. For ICD-10 prediction, ↑ indicates improvement of multi-source training (ICD-10 + ICD-9) compared to training on ICD-10 alone.

4.3 Training and Parameter Efficiency

Mixing multiple ICD sources for training also improves efficiency as the model can now be applied to datasets with different ICD versions, in effect reducing the number of model parameters required. In the context of this study, three PLM-ICD models are needed for the three benchmarks on frequent code prediction, whereas one DUALLAAT can handle them all. Training with mixed data also shortens convergence time: our training plateaued at around 10 epochs compared to 20 epochs for PLM-ICD.

Table 3 reports the comparison between PLM-ICD and DUALLAAT. While DUALLAAT with mixed datasets achieves comparable performance, it uses only a fraction of the training time and parameters. This can be an important factor to consider when deploying a model in a low-resource hospital or outpatient setting (Wu et al., 2022).

	F1 Micro	Precision@8	MAP
MIMIC-IV-ICD10			
Single Dataset	8.4	5.8	19.5
+ Mixing Alt ICD	10.7 (27.4% ↑)	6.1 (5.2% ↑)	21.4 (9.7% ↑)
MIMIC-IV-ICD9			
Single Dataset	7.2	5.6	25.4
+ Mixing Alt ICD	10.9 (51.4% ↑)	6.6 (17.9% ↑)	31.2 (22.8% ↑)

Table 2: Results on the rare codes (frequency < 10 in the cohort). DUALLAAT$_{cnn}$ trained on the corresponding train set was used as baseline.

	Training time	# Param
ICD-10 only		
PLM-ICD	34 hrs	137M
DUALLAAT$_{cnn}$	5 hrs	15M
DUALLAAT	18 hrs	37M
ICD-10 + ICD-9 (×3 datasets)		
PLM-ICD (×3 models)	95 hrs	402M
DUALLAAT$_{cnn}$	17 hrs	15M
DUALLAAT	57 hrs	37M

Table 3: Comparison of training time and model size. Mixed training data includes one dataset for ICD-10 and two datasets for ICD-9.

5 Discussion & Conclusion

Training effective ICD coding models is challenging due to the long-tail distribution of codes and the continuous updates to the ICD system. Most existing work focuses on improving modeling methods, yet little attention has been paid to the impact of the training data itself. We show that mixing multiple data sources – even with varied ICD versions – benefits both frequent and rare code predictions. The gains are notable given that less than 25% of ICD-9 have equivalent mappings to ICD-10 (Fung et al., 2021), suggesting the benefits arise from shared clinical semantics rather than direct label overlap, which warrants future investigation.

The viability of mixing ICD versions means **legacy patient records can be utilized in training new models**, an important implication for developing coding systems to handle ongoing modifica-

	AUC-ROC Micro	AUC-ROC Macro	F1 Micro	F1 Macro	Precision@8
PLM-ICD (Luo et al., 2024)	99.0	91.9	57.0	4.9	69.5
CoRelation (Luo et al., 2024)	99.6	97.2	57.8	6.3	70.0
DUALLAAT (ICD-10 only)	99.6	97.0	56.6	10.2	69.3
DUALLAAT	99.7	97.4	57.1	11.2	69.8

Table 4: Results on 26K ICD-10 codes in MIMIC-IV. PLM-ICD and CoRelation results are from the original paper, which uses a different preprocessing pipeline; we include this approximate comparison for reference.

tions of ICD rules. For major rule updates, a model trained on legacy records can avoid cold-start issues and serve as a foundation for fine-tuning.

In conclusion, we demonstrate the advantages of mixing multi-source datasets to train ICD prediction models, resulting in more accurate, robust, and efficient performance. The synergistic effect of combining diverse sources demonstrates the potential for further scaling both training data and model capacity to advance the progress of ICD coding research. Finally, our code and model checkpoints are released to support reproducibility.[2]

Limitations

ICD and Data Coverage. This study only considered the MIMIC databases as the data sources, which originate from a single institution in the United States and use ICD Clinical Modification (CM). Other regional implementations of ICD exist, such as ICD-10-CA for Canada and ICD-10-AM for Australia. Furthermore, many countries adopt ICD systems in non-English languages, including China (CCD/ICD-10-CN) and Germany (ICD-10-GM). We were limited by the scope of data in this initial study and focused on openly available datasets. Given insights from multilingual language modeling, cross-lingual knowledge transfer (Artetxe et al., 2020) may also be feasible for ICD modeling. However, this warrants future research, particularly to address nuanced differences between different ICD versions beyond mere translation and semantic mappings.

Classification Model. For this study, we employed simple classification models in the experiments, but we believe the patterns of data mixing can be extrapolated to other models given the existing findings on modeling architectures (Ji et al., 2024; Edin et al., 2023). Training PLM-based models with multi-source datasets represents a natural follow-up, though we leave this open as newer and more effective modeling methods may emerge in the near future.

Record Overlap. An inherent challenge in mixing data sources is potential overlapping records. MIMIC-III (2001-2012) and MIMIC-IV (2008-2018) overlap temporally, but we could not explicitly identify shared patients because they use different ID conventions and are not linkable. This may have influenced results for ICD-9, although these were not the primary focus of our experiments. Meanwhile, this is less of an issue for ICD-10 since it is distinctly different from ICD-9 and only exists in MIMIC-IV. We found seven patients appearing in both ICD-9 and ICD-10 datasets from Edin et al. (2023) in MIMIC-IV, with only one patient in the ICD-10 test set (19,802 patients). Given the large test size and different data distributions, we deemed it acceptable to retain this patient to enable direct comparison with the baselines.

[2]The training code and model checkpoints can be found in https://github.com/JHLiu7/Dual-LAAT.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

James C Douglas, Yidong Gan, Ben Hachey, and Jonathan K Kummerfeld. 2025. Less is more: Explainable and efficient ICD code prediction with clin-

ical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2572–2582, New York, NY, USA. Association for Computing Machinery.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. 2021. The clinician and dataset shift in artificial intelligence. *The New England journal of medicine*, 385(3):283–286.

Kin Wah Fung, Julia Xu, Shannon McConnell-Lamptey, Donna Pickett, and Olivier Bodenreider. 2021. Feasibility of replacing the ICD-10-CM with the ICD-11 for morbidity coding: A content analysis. *Journal of the American Medical Informatics Association*, 28(11):2404–2411.

Yidong Gan, Maciej Rybinski, Ben Hachey, and Jonathan K Kummerfeld. 2025. Aligning AI research with the needs of clinical coding workflows: Eight recommendations based on US data analysis and critical review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 909–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, 21(Suppl 6):206.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *An International Journal on Information Fusion*, 118(102963):102963.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. 2024. A unified review of deep learning for automated medical coding. *ACM computing surveys*.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Leah S Larkey and W Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 289–297, New York, NY, USA. Association for Computing Machinery.

Rumeng Li, Xun Wang, and Hong Yu. 2025a. Improving rare and common ICD coding via a multi-agent LLM-based approach. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4945–4949, New York, NY, USA. ACM.

Xiaobo Li, Yijia Zhang, Xiaodi Hou, Shilong Wang, and Hongfei Lin. 2025b. Deep learning for automatic ICD coding: Review, opportunities and challenges. *Artificial intelligence in medicine*, 168(103187):103187.

Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2023. Automated ICD coding using extreme multi-label long text transformer-based models. *Artificial intelligence in medicine*, 144(102662):102662.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. 2024. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007.

Andreas Geert Motzfeldt, Joakim Edin, Casper L Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code like humans: A multi-agent solution for medical coding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22612–22627, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans,

Louisiana. Association for Computational Linguistics.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated ICD coding using deep learning. *arXiv [cs.CL]*.

Ashley Simmons, Kullaya Takkavatakarn, Megan McDougal, Brian Dilcher, Jami Pincavitch, Lukas Meadows, Justin Kauffman, Eyal Klang, Rebecca Wig, Gordon Smith, Ali Soroush, Robert Freeman, Donald J Apakama, Alexander W Charney, Roopa Kohli-Seth, Girish N Nadkarni, and Ankit Sakhuja. 2025. Extracting international classification of diseases codes from clinical documentation using large language models. *Applied clinical informatics*, 16(2):337–344.

Soroush Ali, Glicksberg Benjamin S., Zimlichman Eyal, Barash Yiftach, Freeman Robert, Charney Alexander W., Nadkarni Girish N, and Klang Eyal. 2024. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association: JAMIA*, 17(6):646–651.

Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems*, 30:5998–6008.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T C Poon, Natalie Fitzpatrick, Adam P Levine, Luke T Slater, Alex Handy, Andreas Karwath, Georgios V Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, and 4 others. 2022. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *NPJ digital medicine*, 5(1):186.

John Wu, David Wu, and Jimeng Sun. 2024. Beyond label attention: Transparency in language models for automated medical coding via dictionary learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8848–8871, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. 2023. Surpassing GPT-4 medical coding with a two-stage approach. *arXiv [cs.CL]*.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Jiyang Zheng, Islam Nassar, Thanh Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong, and Yuan-Fang Li. 2025. MedDCR: Learning to design agentic workflows for medical coding. *arXiv [cs.AI]*.

A Datasets for Evaluation

Table 5 presents the statistics of the three benchmark datasets from Edin et al. (2023). We additionally extracted the infrequent codes that were not included in the prior study to evaluate the models' robustness on tail-end ICD codes.

B Classification Model in Detail

B.1 DUALLAAT with dual encoders

Given a clinical note N with t_{note} word tokens $w_1^N, w_2^N, \ldots, w_{t_{\text{note}}}^N$, we have a clinical code C with t_{code} word tokens $w_1^C, w_2^C, \ldots, w_{t_{\text{code}}}^C$ from the code description. They are transformed into word embeddings by a shared embedding layer into $\mathbf{e}_1^N, \mathbf{e}_2^N, \ldots, \mathbf{e}_{t_{\text{note}}}^N$ and $\mathbf{e}_1^C, \mathbf{e}_2^C, \ldots, \mathbf{e}_{t_{\text{code}}}^C$, respectively.

Clinical note representations at token level $\mathbf{H}_{\text{note}} \in \mathbb{R}^{d_{\text{note}} \times t_{\text{note}}}$ are created by a note encoder as $encoder_{note}$:

$$\mathbf{H}_{\text{note}} = \text{encoder}_{\text{note}}(\mathbf{e}_1^N, \ldots, \mathbf{e}_{t_{\text{note}}}^N) \quad (1)$$

The clinical code representation is encoded by a separate $encoder_{code}$ and pooled into a code-level representation. The representation for one code $\mathbf{h}_{\text{code}} \in \mathbb{R}^{d_{\text{code}}}$ is created as:

$$\mathbf{h}_{\text{code}} = \text{Pooling}(\text{encoder}_{\text{code}}(\mathbf{e}_1^C, \ldots, \mathbf{e}_{t_{\text{code}}}^C)) \quad (2)$$

In practice, we consider L codes by encoding them in a batch to create $\mathbf{H}_{\text{code}} \in \mathbb{R}^{|L| \times d_{\text{code}}}$. Then we have two projection matrices $\mathbf{W}_{\text{note}} \in \mathbb{R}^{d_{\text{shared}} \times d_{\text{note}}}$ and $\mathbf{W}_{\text{code}} \in \mathbb{R}^{d_{\text{code}} \times d_{\text{shared}}}$ for note and code, respectively. The projected code representations $\mathbf{H}_{\text{code}} \cdot \mathbf{W}_{\text{code}} \in \mathbb{R}^{|L| \times d_{\text{shared}}}$ replaces the

	MIMIC-III *ICD-9*	MIMIC-IV *ICD-9*	MIMIC-IV *ICD-10*
Number of notes	52,712	209,326	122,278
Notes: Train/val/test [%]	72.9/10.6/16.6	73.8/10.5/15.7	72.9/10.9/16.2
	Frequent Code Set		
Number of unique codes	3,681	6,150	7,942
Codes per note: Median (IQR)	14 (10-20)	12 (8-17)	14 (9-20)
Codes per note: Mean (Std)	15.6±8.0	13.3±7.6	15.7±8.7
	Full Code Set		
Number of unique codes	8,925	11,324	26,085
Codes per note: Median (IQR)	14 (10-20)	12 (8-17)	15 (10-21)
Codes per note: Mean (Std)	15.9±8.1	13.4±7.6	16.0±8.9
	Rare Code Set		
Number of unique codes	5,244	5,174	18,143
Codes per note: Median (IQR)	1 (1-2)	1 (1-1)	1 (1-2)
Codes per note: Mean (Std)	1.5±1.1	1.2±0.6	1.6±1.2
Number of notes (at least 1 rare code)	10,653	14,185	30,332

Table 5: Benchmark datasets on MIMIC-III (ICD-9), MIMIC-IV (ICD-9), and MIMIC-IV (ICD-10) from Edin et al. (2023), which are based on the *frequent code set*. Additional *full code set* and *rare code set* are also examined. Rare codes refer to codes occurring less than 10 times in the dataset, and were not included in Edin et al. (2023). *Full code set* share the same clinical notes with the *frequent code set*. IQR: interquartile range. Std: standard deviation.

code-specific embeddings in typical LWA implementation. With an additional dimension d_{shared} specified, the attention scores and representations based on dual encoders are computed:

$$\mathbf{A}_{\text{dual}} = \text{softmax}\Big(\tanh(\mathbf{H}_{\text{code}} \cdot \mathbf{W}_{\text{code}}) \cdot \tanh(\mathbf{W}_{\text{note}} \cdot \mathbf{H}_{\text{note}}) \Big) \quad (3)$$

$$\mathbf{J}_{\text{dual}} = \mathbf{H}_{\text{note}} \cdot \mathbf{A}_{\text{dual}}^{\top} \quad (4)$$

This formulates the backbone of DUALLAAT, which use separate note encoder and code encoder to create their corresponding representations and compute label-wise attention.

To further enhance the dual label-wise attention, we also consider M attention heads with $\mathbf{W}^1_{\text{note}}, \ldots, \mathbf{W}^M_{\text{note}}$ and $\mathbf{W}^1_{\text{code}}, \ldots, \mathbf{W}^M_{\text{code}}$ as multi-head attention (MHA) (Vaswani et al., 2017). Then we compute $\mathbf{J}^1_{\text{dual}}, \ldots, \mathbf{J}^M_{\text{dual}}$ and obtain the multi-head note representation per code by concatenation:

$$\mathbf{J}^{\text{mha}}_{\text{dual}} = \text{concat}(\mathbf{J}^1_{\text{dual}}, \ldots, \mathbf{J}^M_{\text{dual}}) \quad (5)$$

The final output is computed as:

$$\hat{\mathbf{y}} = \text{sigmoid}(\text{classifier}(\mathbf{J}^{\text{mha}}_{\text{dual}})) \quad (6)$$

The model is trained to minimize binary cross entropy loss. Different from the previous models that consider only note N as input, DUALLAAT now takes both N (notes) and C (codes) as inputs. This enables a flexible choice of L by changing C either during training or at inference. Since the clinical codes C are provided as code descriptions in text, it allows flexibility to feed different versions of ICD codes as long as they are provided in the textual form. During inference, these descriptions can be directly mapped back to the actual codes.

B.2 Training DUALLAAT

To train DUALLAAT in batches, we need to dynamically set the label space L for different codes. For example, given a batch of 32 patients with each patient having 20 codes on average, L would be set dynamically according to appearing clinical codes with $|L| = 640$ (20×32). Meanwhile, in reality the number of clinical codes can vary dramatically from patient to patient, resulting in large variations in code numbers that would destabilize the label space and thus training. In addition, patients in a batch sometimes may contain codes in overlap, in effective reducing the scope of L.

To address this, we specify the label space L to be a fixed number that is larger than the possible codes to appear in a batch. Given the fixed L, we collect the unique codes in the batch as positives L_{pos}, and then randomly sample negative codes L_{neg} from a code pool to form L so that $|L| = |L_{pos}| + |L_{neg}|$. This setup has three advantages. First, it fixes the number of L to stabilize training. Second, it removes duplicated positive codes and allows negative sampling to expand the scope of L, enabling higher sample efficiency. Finally, $|L|$ becomes a hyperparameter to be controlled.

During training, we use a sampler to keep only

one ICD version in each batch, i.e., L is either L^{icd9} or L^{icd10}. These batches of different ICD versions are mixed and shuffled to train the model with standard mini-batch gradient descent.

B.3 Hyperparameter setting

We set the dimensions of the encoders and the projection layers the same $d_{note} = d_{code} = d_{shared}$ without further tuning. We follow the implementation from (Edin et al., 2023) as best practice in setting the hyperparameters, where DUALLAAT has a dimension of 512, one bidirectional layer, and dropout rate of 0.3; and DUALLAAT$_{\text{cnn}}$ has a 256 filters with filter width of 10, and dropout rate of 0.2. All models are trained using a linear scheduler with learning rate of 0.001 and 2000 warmup steps on batches of 32 samples. DUALLAAT is additionally imposed with weight decay of 0.001 to regularize training. The maximum number of tokens in clinical notes is set to 4000, and that for the clinical code descriptions is set to 48.

We pretrain the word embeddings (100 dimensions) based on the three train sets, again following the setup in (Edin et al., 2023). By default, we train the models on full ICD codes, and set the label space size $|L|$ to 8192. We use GRU as the default RNN choice for DUALLAAT$_{\text{rnn}}$, with an alternative DUALLAAT$_{\text{cnn}}$ using CNN.

	CAML	LAAT	PLM-ICD	DUALLAAT$_{cnn}$ (+ ICD-9)	DUALLAAT (+ ICD-9)
AUC-ROC Micro	98.5 ± 0.0	99.0 ± 0.1	99.2 ± 0.0	99.3 ± 0.0	$\mathbf{99.3 \pm 0.0}$
AUC-ROC Macro	91.1 ± 0.1	95.4 ± 0.3	96.6 ± 0.2	97.2 ± 0.0	$\mathbf{97.4 \pm 0.0}$
F1 Micro	55.4 ± 0.2	57.9 ± 0.1	$\mathbf{58.5 \pm 0.7}$	56.6 ± 0.2	58.0 ± 0.0
F1 Macro	16.0 ± 0.3	20.3 ± 0.4	21.1 ± 2.3	26.0 ± 0.2	$\mathbf{29.9 \pm 0.3}$
P@8	66.8 ± 0.2	68.9 ± 0.1	$\mathbf{69.9 \pm 0.6}$	68.4 ± 0.2	$\mathbf{69.9 \pm 0.1}$
P@15	52.2 ± 0.1	54.3 ± 0.1	$\mathbf{55.0 \pm 0.6}$	53.6 ± 0.2	54.9 ± 0.1
R-precision	54.5 ± 0.2	57.2 ± 0.1	57.9 ± 0.8	56.4 ± 0.2	$\mathbf{58.0 \pm 0.1}$
MAP	57.4 ± 0.2	60.6 ± 0.2	61.9 ± 0.9	60.4 ± 0.2	$\mathbf{62.3 \pm 0.1}$

Table 6: Results on frequent ICD-10 codes (main benchmark) with additional baseline models. Scores are reported with standard deviation.

EmCellLLM: Human Peri-Implantation Embryonic Cell Annotation Based on Large Language Models

Xiaorui Guo[1] **Zhiwei Liu**[2] **Qianqian Xie**[3,4] **Sophia Ananiadou**[2,5]
[1]The University of Edinburgh [2]The University of Manchester
[3]School of Artificial Intelligence, Wuhan University
[4]Center for Language and Information Research, Wuhan University [5]ELLIS Manchester
X.Guo-46@sms.ed.ac.uk, {zhiwei.liu,sophia.ananiadou}@manchester.ac.uk
xieq@whu.edu.cn

Abstract

The advent of single-cell RNA sequencing has enabled unprecedented resolution of cell fate decisions and regulatory mechanisms during peri-implantation human embryogenesis, in which accurate cell type annotation is a fundamental prerequisite and the first step for subsequent fate and mechanism inference. Large language models (LLMs) have demonstrated outstanding performance in various fields. However, current studies mostly rely on traditional methods and have not explored the application of LLMs in the field of human embryonic cell annotation. The main reason is the lack of instruction tuning datasets and evaluation benchmarks. In this paper, we proposed EmCellLLM, the first open sourced LLMs that are specialized for human embryonic cell type prediction task based on fine-tuning Qwen3-8B with EmCell4Instruction, the first embryonic cell type prediction instruction dataset. To support LLM instruction tuning, we also build EmCellBench, the first benchmark for evaluating human embryonic cell type prediction ability of LLMs. We compare our models with a variety of LLMs on EmCellBench, where our model outperforms all other open-sourced LLMs as well as DeepSeek.

1 Introduction

Human peri-implantation embryos (approximately days 5–14) represent a critical developmental window in which the embryo transitions from an undifferentiated state to distinct lineages while interacting with the endometrium (Zhai et al., 2022; David et al., 2023). Abnormal development during this period is strongly associated with implantation failure, early pregnancy loss, placental dysfunction, and developmental origins of disease (Norwitz et al., 2001).

Performing single-cell sequencing on the peri-implantation stage and accurately annotating the cells allows the resolution of lineage trajectories, dynamic transitional states, and key regulatory networks at single-cell resolution, thereby reconstructing a comprehensive map of early human development (Ostrer et al., 2006; Latham, 2023; Anger et al., 2025). The challenges of single-cell annotation in peri-implantation human embryos stem from three main factors. (1) Human embryo samples are scarce due to ethical and practical limitations, and each developmental stage contains only a small number of cells, resulting in sparse and noisy datasets. (Petropoulos et al., 2016; Gul and Zhang, 2025). (2) This stage is highly dynamic, with rapid cell fate diversification and continuous, making clear cell-type boundaries and marker-based annotation difficult (Molè et al., 2021). (3) Reference atlases are usually constructed by integrating datasets from multiple sources, but strong batch effects arising from different experimental protocols and sequencing platforms can obscure biological signals and lead to inconsistent or inaccurate annotation (Weatherbee et al., 2023). Accurate annotation is therefore essential for understanding early human development and its applications in reproductive medicine and stem cell modeling.

Traditional embryonic cell annotation relies on known marker genes and manual expert judgment to identify cell types and manually assign cluster labels which is time-consuming, hard to scale, and limited by incomplete marker knowledge and subjective interpretation (Luecken and Theis, 2019; Cheng et al., 2023). Automated cell annotation methods reduce manual curation by leveraging marker-based approaches (e.g., CellMarker 2.0), reference mapping and label transfer methods (e.g., Seurat, SingleR, scmap, TemporalVAE), and supervised or deep learning classifiers (e.g., CellTypist, scTab, scANVI/scVI) (Hu et al., 2023; Stuart et al., 2019; Aran et al., 2019; Kiselev et al., 2018; Liu et al., 2025b; Domínguez Conde et al., 2022; Fischer et al., 2024; Lopez et al., 2018; Xu et al., 2021a). These methods assign cell identities by

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 382–391
July 3-4, 2026 ©2026 Association for Computational Linguistics

mapping query data to reference atlases or learning expression–label relationships. While effective for well-characterized cell types, their performance depends on reference coverage and annotation consistency and is sensitive to batch effects and cross-dataset variation (Cheng et al., 2023; Zhao et al., 2025). In human embryogenesis, where cell states are continuous and reference atlases integrated from multiple sources and incomplete, these methods often fail to resolve rare populations, leading to misannotation or low-confidence predictions.

LLM-based automated cell annotation methods have advanced rapidly in recent years, with the core idea of transforming single-cell transcriptomic data into representations that can be processed by LLMs (Hou and Ji, 2024). For example, the C2S approach converts single-cell gene expression profiles into text sequences by ranking genes in descending order of expression levels, enabling large language models to directly "read" gene expression patterns. While preserving the complexity of single-cell data, this formulation generates biologically meaningful cell representations that can be used for downstream cell type prediction (Levine et al., 2024).

To address the above issues, we construct the first instruction-tuning dataset for embryonic cell type prediction, EmCell4Instruction, to support LLM fine-tuning, which includes a single-cell transcriptome atlas integrating eight human embryo datasets spanning developmental stages from day 3 to day 18.5. Based on this dataset, we propose EmCellLLM, the first open-source LLM for embryonic cell type prediction, which developed by fine-tuning Qwen3-8B on EmCell4Instruction. To systematically evaluate model performance, we further introduce EmCellBench, a benchmark for embryonic cell type annotation. Experimental results on EmCellBench demonstrate that EmCellLLM achieves state-of-the-art performance among open-source LLMs, and we will release both the dataset and the model to the public.

Our main contributions are as follows:

(1) We construct EmCell4Instruction, the first embryonic cell type prediction instruction-tuning dataset.

(2) We develop EmCellLLM, the first open-sourced LLMs that are specialized for the embryonic cell type prediction task.

(3) We build EmCellBench, the first benchmark to evaluate the embryonic cell type prediction ability of LLMs. The results on EmCellBench demonstrate that our model overtakes other open-sourced LLMs.

2 Related work

2.1 Cell Type Annotation

There are few studies semi-automatic or automatic annotate human embryonic cells. Most of them are based on traditional label transfer methods or deep learning methods. Xu et al. (2021b) use Seurat-based integration combined with marker gene curation and reference-guided label transfer for semi-automatic annotation. Zhao et al. (2025) use Seurat-based integration and label transfer to project query datasets onto a reference human embryo atlas for cell type annotation. TemporalVAE annotates cell types by mapping new cells into an atlas-informed latent space and transferring labels from nearest neighbors (Liu et al., 2025b). (Proks et al., 2025) use deep learning models to integrate single-cell RNA-seq data from mouse and human pre-implantation embryos for cross-species cell type annotation and developmental state prediction. However, there is currently no open-sourced LLM specifically designed for the annotation of human embryonic cells.

2.2 Open Sourced Large Language Models

Significant research efforts have focused on developing open-source LLMs as alternatives to closed-source models such as GPT series, with the aim of enabling more accessible research on improving and applying LLMs. Well-known series of open-sourced, general-purpose language models include LLama series (Dubey et al., 2024), DeepSeek series (Liu et al., 2025a), Qwen series (Yang et al., 2025). There are also many open-sourced LLMs for specific domains, including FMDLlama (Liu et al., 2025d) for financial misinformation detection, human cell C2S LLMs (Levine et al., 2024) for somatic cell analysis, ExTES-LLaMA (Zheng et al., 2023) for emotional support chatbots, MeLLaMA for Medical Applications (Xie et al., 2024), and ConspEmoLLM (Liu et al., 2024, 2025c) for conspiracy detection. In this work, we extend the inventory of domain-specific LLMs, by developing the first open-sourced LLM for embryonic cell type annotation.

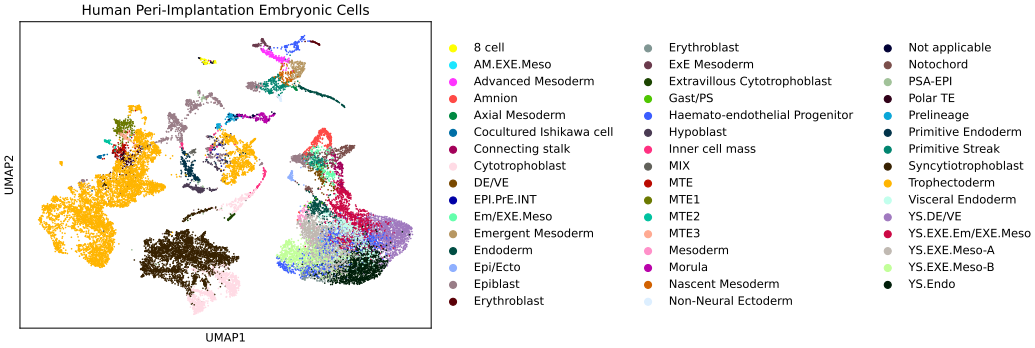


Figure 1: UMAP of Human Peri-Implantation Embryonic Cells

3 Methods

3.1 Task formalization

We approach cell type detection as a generative task, applying a generative model as a foundation. This generative model is an autoregressive language model $P_\phi(y|x)$, parameterized using pre-trained weights ϕ. Each task (t) is represented as a set of context-target pairs: $D_t = (q_i^t, r_i^t)_{i=1,2,...N_t}$, where the context q is a token sequence containing the task description, cell sentence, and query, and r is a further token sequence containing the answer to the query. The model is optimized based on the training dataset, with the aim of maximizing the objective of conditional language modeling to improve prediction and generation performance.

3.2 Construction of instruction tuning dataset

Data	Train	Test
Embryonic cells	19652	4913

Table 1: Dataset statistics. Detailed statistics of each cell type can be found at Table 3.

3.2.1 Raw data

We build our instruction tuning dataset based on a single-cell human embryonic cells transcriptome atlas which integrates 8 human embryo datasets and spanning from embryonic development day 3 to day 18.5 (Liu et al., 2025b). Among them, data from Xiao et al. (2024), Cui et al. (2025) and Tyser et al. (2021) all come from one in vivo human embryo, data from Petropoulos et al. (2016) and Molè et al. (2021) come from 88 and 16 ex vivo embryos, and data from Xiang et al. (2020), Zhou et al. (2019) and Liu et al. (2022) come from 42, 65 and 25 in vitro embryos, individually. Figure 1 presents the UMAP of the datasets.

3.2.2 Construction of the EmCell instruction tuning dataset (EmCell4Instruction) and EmCell benchmark (EmCellBench)

We use the raw dataset without batch correction as the basis for building the instruction dataset. We split the original data into train and test sets. The dataset statistics are presented in Table 1. We construct instruction-tuning data based on the templates in Table 4. One of the 15 instruction templates is randomly selected for each item to enhance diversity. *[num_genes]* represents the top K highly expressed genes used for cell type classification. In this paper, we set K to 200. *[cell_sentence]* is obtained by converting cell expression profiles into natural language sentences in C2S (Levine et al., 2024), which transforms single-cell gene expression data into rank-ordered gene sequences by expression level, enabling LLMs to effectively process transcriptomic information while leveraging standard transformer libraries. LLMs directly output the predicted cell type as their response. Figure 2 presents examples used to fine-tune the LLM.

After constructing the instruction data. We collect the train data as instruction-tuning data (EmCell4Instruction) and test data as the embryonic cells type prediction benchmark (EmCellBench), which are used to fine-tune the LLMs and evaluate the ability of LLMs in the embryonic cells type prediction domain, respectively.

3.3 EmCellLLM

We built EmCellLLM by fine-tuning Qwen3-8b (Yang et al., 2025) using the EmCell4Instruction dataset. The models are trained based on the AdamW optimizer (Loshchilov and Hutter, 2017) for five epochs, using DeepSpeed (Rasley et al., 2020) to reduce memory usage. We set the batch size to 128. The initial learning rate is set to 1e-5 with a warm-up ratio of 0.1. All models are trained

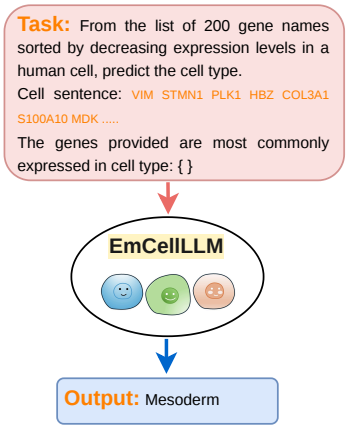


Figure 2: An overview of instruction tuning of EmCellLLM

Model	ACC	PRE	REC	F1
Qwen3-8b-R	0.135	0.023	0.029	0.008
Qwen3-14b-R	0.161	0.053	0.056	0.028
Qwen3-32b-R	0.148	0.029	0.035	0.014
Qwen3-8b	0.138	0.018	0.030	0.007
Qwen3-14b	0.129	0.027	0.045	0.012
Qwen3-32b	0.120	0.034	0.046	0.019
Qwen2.5-72b	0.078	0.004	0.005	0.002
Llama3.1-8b	0.062	0.015	0.025	0.009
Llama3.3-70b	0.092	0.039	0.063	0.026
DeepSeek	0.168	0.068	0.052	0.032
C2S-Pythia-410M	0.884	0.770	0.752	0.745
EmCellLLM	**0.905**	**0.858**	**0.811**	**0.827**

Table 2: Results on EmCellBench.

on four NVIDIA L40s GPUs, each with 48GB of memory. We compare with the domain-specific C2S-Pythia-410M (Levine et al., 2024) cell type prediction model, which has been trained on multiple datasets from CellxGene and the Human Cell Atlas for cell type prediction. We further continue training based on EmCell4Instruction as the baseline. Figure 2 provides an overview of instruction tuning of EmCellLLM for the embryonic cell type prediction task.

4 Experiments

4.1 Baseline models

Our extensive evaluation of open-source and proprietary LLMs included the following reasoning-focused models: Qwen3 reasoning variants (8B-R, 14B-R, and 32B-R) (Yang et al., 2025). We also assessed a diverse set of no-reasoning LLMs, DeepSeek-V3.2-Chat (DeepSeek-C) (Liu et al., 2025a), the Qwen3 no-reasoning models (8B, 14B, and 32B), Qwen2.5-72B-Instruct (Qwen72B) (Qwen et al., 2025), Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct (Dubey et al., 2024). Additionally, we compare with the domain-specific C2S-Pythia-410M cell type prediction model, which has been trained on multiple datasets from CellxGene and the Human Cell Atlas for cell type prediction. We further continue training based on EmCell4Instruction as the baseline.

4.2 Evaluation methods

We uses metrics such as Accuracy, Precision, Recall, Macro-F1 for embryonic cell type detection task.

4.3 Results

Table 2 presents the results on EmCellBench. From the table, we can see EmCellLLM achieve SOTA results among all other open-sourced LLMs, including DeepSeek. C2S-Pythia-410M follows closely, demonstrating the effectiveness of the instruction-tuning approach. Other open-source general-purpose LLMs perform poorly, likely because they lack domain-specific understanding and require further training to acquire such knowledge. Reasoning models of the same size perform slightly better than their non reasoning counterparts such as the Qwen3 series, indicating that reasoning provides an advantage when handling unseen and complex knowledge tasks. Larger models generally achieve better performance, and among the untrained models, DeepSeek performs the best, suggesting that models with more parameters tend to possess richer knowledge and stronger overall capabilities.

5 Conclusion

In this paper, we propose EmCellLLM, the first LLM for embryonic cell annotation. We also construct an embryonic cell prediction instruction dataset (EmCell4Instruction) and an evaluation benchmark (EmCellBench). We conduct a comprehensive analysis of the performance of EmCellLLM, as well as a variety of LLMs on the EmCellBench benchmark. The results indicate that EmCellBench performs exceptionally well in the embryonic cell annotation task, achieving SOTA compared to the other open-sourced LLMs.

In the future, we aim to augment the EmCell4Instruction and EmCellBench datasets by collecting a more comprehensive single-cell sequencing atlas of human embryonic development, which can help further improve the EmCellLLM and evaluate the embryonic cell annotation ability of LLMs more comprehensively.

Limitations

The potential limitations of our work may be summarized as follows:

Due to restricted computational resources, we only carried out instruction-tuning/evaluation of the task using 8b/14b/32b/70b/72b LLMs. As such, we have not considered the impact of using larger models on the Embryonic cell type prediction task.

Due to the scarcity and limited temporal coverage of in vivo human embryo data, we integrate ex vivo and in vitro datasets to increase diversity and scale, but these remain less ideal than in vivo data due to potential domain shifts and batch effects.

While our dataset covers embryonic stages from day 3 to day 18.5, spanning the peri-implantation period, extending it to include the full developmental trajectory from day 0 to birth would enable the construction of a more comprehensive human embryonic atlas and improve the robustness and generalizability of annotation models.

Acknowledgments

This work was supported by the British Heart Foundation Manchester Research Excellence Award (RE/24/130017). This work was supported by the CCF-Tencent Rhino-Bird Open Research Fund (CCF-Tencent RAGR20250115), Wuhan Natural Science Foundation Exploratory Program (Morning Light Program) Project (2026040301020029) and the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan.

References

Martin Anger, Katerina Komrskova, Ahmed Z Balboula, Paolo Rinaudo, and Jason G Knott. 2025. Fertilization and early embryogenesis: from research to clinical practice. *Frontiers in Cell and Developmental Biology*, 12:1547205.

Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172.

Changde Cheng, Wenan Chen, Hongjian Jin, and Xiang Chen. 2023. A review of single-cell rna-seq annotation, integration, and cell–cell communication. *Cells*, 12(15):1970.

Lina Cui, Sirui Lin, Xiaolong Yang, Xinwei Xie, Xiaoyan Wang, Nannan He, Jingyu Yang, Xin Zhang, Xiaojian Lu, Xiaodi Yan, et al. 2025. Spatial transcriptomic characterization of a carnegie stage 7 human embryo. *Nature cell biology*, 27(2):360–369.

Laurent David, Alexandre Bruneau, Thomas Fréour, Nicolas Rivron, and Hilde Van de Velde. 2023. An update on human pre-and peri-implantation development: a blueprint for blastoids. *Current opinion in genetics & development*, 83:102125.

César Domínguez Conde, Chao Xu, Louie B Jarvis, Daniel B Rainbow, Sara B Wells, Tamir Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. 2022. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Felix Fischer, David S Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J Theis. 2024. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611.

Samina Gul and Chengting Zhang. 2025. Single cell rna sequencing and its impact on understanding human embryo development. *International Journal of Molecular Sciences*, 26(16):7741.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature methods*, 21(8):1462–1465.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. 2023. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic acids research*, 51(D1):D870–D876.

Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. 2018. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362.

Keith E Latham. 2023. Preimplantation embryo gene expression: 56 years of discovery, and counting. *Molecular reproduction and development*, 90(4):169–200.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. 2024. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pages 2023–09.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Dandan Liu, Yidong Chen, Yixin Ren, Peng Yuan, Nan Wang, Qiang Liu, Cen Yang, Zhiqiang Yan, Ming Yang, Jing Wang, et al. 2022. Primary specification of blastocyst trophectoderm by scrna-seq: New insights into embryo implantation. *Science Advances*, 8(31):eabj3725.

Yijun Liu, Fangxin Cai, Melania Barile, Yi Chang, Dandan Cao, and Yuanhua Huang. 2025b. Temporalvae: atlas-assisted temporal mapping of time-series single-cell transcriptomes during embryogenesis. *Nature Cell Biology*, pages 1–11.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, and Sophia Ananiadou. 2025c. Conspemollm-v2: A robust and stable model to detect sentiment-transformed conspiracy theories. *arXiv preprint arXiv:2505.14917*.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2025d. Fmdllama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Malte D Luecken and Fabian J Theis. 2019. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):MSB188746.

Matteo A Molè, Tim HH Coorens, Marta N Shahbazi, Antonia Weberling, Bailey AT Weatherbee, Carlos W Gantner, Carmen Sancho-Serra, Lucy Richardson, Abbie Drinkwater, Najma Syed, et al. 2021. A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nature communications*, 12(1):3679.

Errol R Norwitz, Danny J Schust, and Susan J Fisher. 2001. Implantation and the survival of early pregnancy. *New England Journal of Medicine*, 345(19):1400–1408.

Harry Ostrer, DI Wilson, and NA Hanley. 2006. Human embryo and early fetus research. *Clinical genetics*, 70(2):98–107.

Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026.

Martin Proks, Nazmus Salehin, and Joshua M Brickman. 2025. Deep learning-based models for preimplantation mouse and human embryos based on single-cell rna sequencing. *Nature methods*, 22(1):207–216.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902.

Richard CV Tyser, Elmir Mahammadov, Shota Nakanoh, Ludovic Vallier, Antonio Scialdone, and Shankar Srinivas. 2021. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, 600(7888):285–289.

Bailey AT Weatherbee, Carlos W Gantner, Lisa K Iwamoto-Stohl, Riza M Daza, Nobuhiko Hamazaki, Jay Shendure, and Magdalena Zernicka-Goetz. 2023. Pluripotent stem cell-derived model of the post-implantation human embryo. *Nature*, 622(7983):584–593.

Lifeng Xiang, Yu Yin, Yun Zheng, Yanping Ma, Yonggang Li, Zhigang Zhao, Junqiang Guo, Zongyong Ai, Yuyu Niu, Kui Duan, et al. 2020. A developmental landscape of 3d-cultured human pre-gastrulation embryos. *Nature*, 577(7791):537–542.

Zhenyu Xiao, Lina Cui, Yang Yuan, Nannan He, Xinwei Xie, Sirui Lin, Xiaolong Yang, Xin Zhang, Peifu Shi, Zhifeng Wei, et al. 2024. 3d reconstruction of a gastrulating human embryo. *Cell*, 187(11):2855–2874.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. 2021a. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):MSB20209620.

Yichi Xu, Tengjiao Zhang, Qin Zhou, Mengzhu Hu, Yao Qi, Yifang Xue, Lihui Wang, Yuxiao Nie, Zhirong Bao, and Weiyang Shi. 2021b. A single-cell transcriptome atlas of human early embryogenesis. *BioRxiv*, pages 2021–11.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinglei Zhai, Zhenyu Xiao, Yiming Wang, and Hongmei Wang. 2022. Human embryonic development: from peri-implantation to gastrulation. *Trends in cell biology*, 32(1):18–29.

Cheng Zhao, Alvaro Plaza Reyes, John Paul Schell, Jere Weltner, Nicolás M Ortega, Yi Zheng, Åsa K Björklund, Laura Baqué-Vidal, Joonas Sokka, Ras Trokovic, et al. 2025. A comprehensive human embryo reference tool using single-cell rna-sequencing data. *Nature methods*, 22(1):193–206.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Fan Zhou, Rui Wang, Peng Yuan, Yixin Ren, Yunuo Mao, Rong Li, Ying Lian, Junsheng Li, Lu Wen, Liying Yan, et al. 2019. Reconstituting the transcriptome and dna methylome landscapes of human implantation. *Nature*, 572(7771):660–664.

A Data Details

Table 3 presents the statistics of each cell type. Table 4 shows the instruction templates for EmCell4Instruction and EmCellBench.

B Error Analysis

Figure 3 presents the confusion matrix of EmCell-LLM on EmCellBench. The confusion matrix reveals the misclassification among several cell types. For example, Gast/PS cells are predominantly predicted as Endoderm. This pattern is likely because of the extremely small training sample size (only 4 cells) prevents the model from learning robust class-specific features. Additionally, EPI.PrE.INT cells are dispersed across multiple predicted classes, including Epiblast, Primitive Endoderm, and Trophectoderm. EPI.PrE.INT cells is labeled as a mix cell population in training data which including Epiblast, Primitive Endoderm, and intermediate cells. Therefore, it's it is reasonable to predict Epiblast as Primitive Endoderm. However, as there are only 12 cells used to train model, model did not receive sufficient information for training causing a part of cells were mislabeled as Trophectoderm.

Cell Type	Full Name	Train	Test
8 cell	8-cell stage embryo	50	14
AM.EXE.Meso	Amnion-associated Extraembryonic Mesoderm	1	35
Advanced Mesoderm	Advanced Mesoderm	129	0
Amnion	Amnion	239	60
Axial Mesoderm	Axial Mesoderm	18	5
Cocultured Ishikawa cell	Cocultured Ishikawa cells	31	12
Connecting stalk	Connecting stalk	160	34
Cytotrophoblast	Cytotrophoblast	868	216
DE/VE	Definitive Endoderm / Visceral Endoderm	106	21
EPI.PrE.INT	Epiblast–Primitive Endoderm Intermediate	12	4
Em/EXE.Meso	Embryonic / Extraembryonic Mesoderm	137	45
Emergent Mesoderm	Emergent Mesoderm	139	46
Endoderm	Endoderm	290	77
Epi/Ecto	Epiblast / Ectoderm	54	17
Epiblast	Epiblast	956	228
Erythroblast	Erythroblast	29	3
Erythroblast	Erythroblast	226	64
ExE Mesoderm	Extraembryonic Mesoderm	70	13
Extravillous Cytotrophoblast	Extravillous Cytotrophoblast	37	3
Gast/PS	Gastrulation / Primitive Streak	4	2
Haemato-endothelial Progenitor	Haemato-endothelial progenitor	995	265
Hypoblast	Hypoblast	171	31
Inner cell mass	Inner Cell Mass	81	15
MIX	Mixed population	31	8
MTE	Mural Trophectoderm	111	23
MTE1	Mural Trophectoderm subtype 1	229	46
MTE2	Mural Trophectoderm subtype 2	35	4
MTE3	Mural Trophectoderm subtype 3	59	12
Mesoderm	Mesoderm	52	12
Morula	Morula	132	29
Nascent Mesoderm	Nascent Mesoderm	74	24
Non-Neural Ectoderm	Non-neural ectoderm	21	8
Not applicable	Not applicable	75	21
Notochord	Notochord	213	68
PSA-EPI	Post-implantation Stage Epiblast	37	7
Polar TE	Polar Trophectoderm	164	51
Prelineage	Pre-lineage stage	86	20
Primitive Endoderm	Primitive Endoderm	181	65
Primitive Streak	Primitive Streak	293	60
Syncytiotrophoblast	Syncytiotrophoblast	2600	655
Trophectoderm	Trophectoderm	5085	1248
Visceral Endoderm	Visceral Endoderm	378	76
YS.DE/VE	Yolk Sac Definitive Endoderm / Visceral Endoderm	1156	242
YS.EXE.Em/EXE.Meso	Yolk Sac Embryonic / Extraembryonic Mesoderm	995	242
YS.EXE.Meso-A	Yolk Sac Extraembryonic Mesoderm subtype A	1151	303
YS.EXE.Meso-B	Yolk Sac Extraembryonic Mesoderm subtype B	615	169
YS.Endo	Yolk Sac Endoderm	1076	310
Total		19652	4913

Table 3: Detailed statistics of each cell type.

Instruction Template
(1) The following is a list of {num_genes} gene names ordered by descending expression level in a human cell. Your task is to give the cell type which this cell belongs to based on its gene expression. Cell sentence: {cell_sentence}. The cell type corresponding to these genes is:
(2) Below is a list of {num_genes} gene names in order of descending expression level from a human cell. Based on this, predict what the cell type of this cell is. Cell sentence: {cell_sentence}. These genes are most likely associated with cell type:
(3) Given the list of {num_genes} gene names ordered by descending expression level from a human cell, identify the cell type. Cell sentence: {cell_sentence}. The probable cell type for these genes is:
(4) Analyze the following list of {num_genes} genes sorted by decreasing expression levels in a human cell and determine its cell type. Cell sentence: {cell_sentence}. Based on these genes, the corresponding cell type is:
(5) From the list of {num_genes} gene names ordered by decreasing expression level in a human cell, infer the cell type of the cell. Cell sentence: {cell_sentence}. These genes suggest the cell type is most likely:
(6) The {num_genes} gene names below are listed by descending expression level in a human cell. Predict the cell type based on this information. Cell sentence: {cell_sentence}. These genes are indicative of cell type:
(7) Below is a list of {num_genes} gene names sorted by descending expression in a human cell. Determine the cell type of this cell from its expressed genes. Cell sentence: {cell_sentence}. The associated cell type for these genes appears to be:
(8) The following list contains {num_genes} gene names ordered by descending expression level in a human cell. Deduce the cell type based on this. Cell sentence: {cell_sentence}. These genes typically correspond to cell type:
(9) Here is a list of {num_genes} genes in order of descending expression level from a human cell. Identify the cell type of this cell based on this information. Cell sentence: {cell_sentence}. The expected cell type based on these genes is:
(10) Examine the list of {num_genes} gene names sorted by decreasing expression in a human cell, and from this, identify its cell type. Cell sentence: {cell_sentence}. These genes are commonly found in cell type:
(11) The {num_genes} gene names below are arranged by descending expression level in a human cell. Determine the cell type of this cell. Cell sentence: {cell_sentence}. The cell type that these genes are most commonly linked with is:
(12) Here is a list of {num_genes} genes ordered by expression level in a human cell. Based on this information, identify the cell type. Cell sentence: {cell_sentence}. Based on the expression levels, the cell type would likely be:
(13) From the list of {num_genes} gene names sorted by decreasing expression levels in a human cell, predict the cell type. Cell sentence: {cell_sentence}. The genes provided are most commonly expressed in cell type:
(14) Below is a list of {num_genes} genes ordered by descending expression in a human cell. Use this information to determine the cell type. Cell sentence: {cell_sentence}. These genes suggest the cell type is most likely:
(15) The {num_genes} gene names below are listed by descending expression levels in a human cell. Based on this, predict the cell type. Cell sentence: {cell_sentence}. The cell type corresponding to these genes is:

Table 4: Instructions used for EmCell4Instruction and EmCellBench.

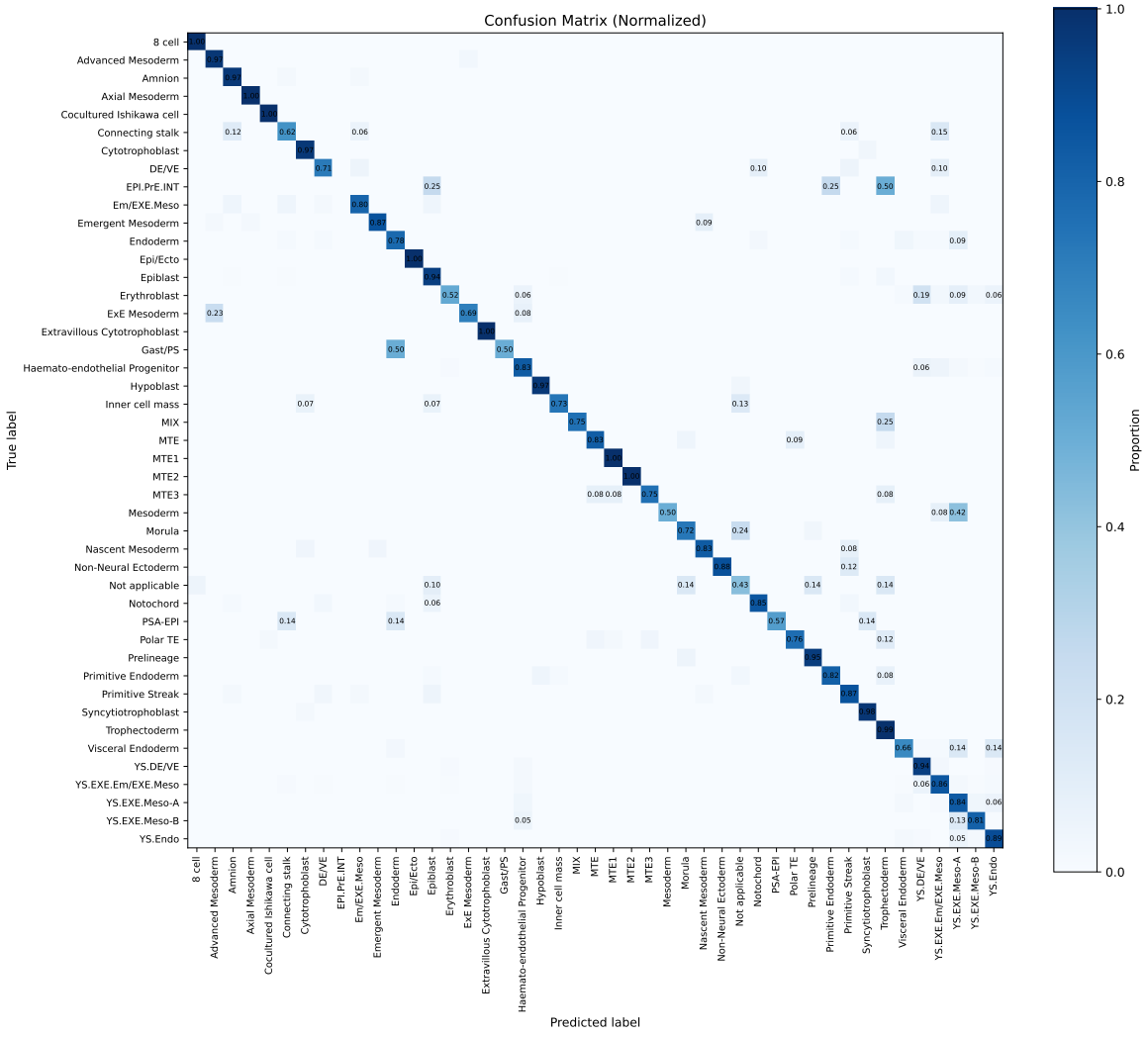


Figure 3: Confusion matrix of EmCellLLM on EmCellBench.

Randomized Controlled Trials as the Gold-Standard for Evaluating LLMs: A Primer for Biomedical NLP Researchers

Vicente Ivan Sanchez Carmona and Shanshan Jiang and Bin Dong
Ricoh Software Research Center (Beijing) Co., Ltd
{Vicente.Carmona, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com

Abstract

Large Language Models (LLMs) are no longer mere laboratory objects of study. LLMs have become everyday tools in society across diverse populations and domains. In clinical contexts, LLMs have already been devised as clinical support applications. However, along with benefits, negative or adverse effects might arise, such as LLMs potentially providing psychologically distressing advice to adolescents when used for mental health support. This raises questions on the benefits of LLMs and calls for real-world evaluations: Are LLMs really helpful and effective for the intended purposes people are using them or will use them for? To answer this type of question we propose to use Randomized Controlled Trials (RCTs). RCTs are considered the most strict experimental design in the fields of Medicine, Psychiatry, Psychology, among others; however, the use of RCTs in the NLP field is almost negligible. In spite of the NLP field being the de facto locus of research on LLMs, other fields, prominently Medicine, are leading the RCT evaluations on LLMs. In this primer paper, we present a concise introduction to the principles of RCTs to guide NLP researchers to design RCT studies for evaluating LLMs.

1 Introduction

In a recent submission to the journal of *Computational Linguistics*, Reiter (2025) argues that the ACL community places a small emphasis on real-world evaluations of NLP applications. While we applaud and echo this argument, we acknowledge that, to the detriment of both the ACL community and the final users, lacking real-world evaluations not only limits our understanding of the efficacy and utility of these applications on those users, it can lead to adverse scenarios, such as users ingesting chemicals with potentially toxic properties as advised by an LLM (Eichenberger et al., 2025).

Critically, evaluation deficits of NLP applications in clinical contexts may potentially open the possibility for negative scenarios such as adolescents being psychologically affected after interactions with LLMs, chatbots, or conversational agents since, according to a survey, it has been estimated that approximately one in eight adolescents and young adults have engaged with online LLMs for mental health advice and to alleviate mental distress (McBain et al., 2025). Although previous works, mainly from clinical fields, have already assessed the impact of generative systems on this outcome, the overall evidence is inconclusive (Feng et al., 2025). In addition, online LLMs are being adopted to other use cases. For example, students of different ages access LLMs for requesting help with homeworks or solving complex problems (Gasaymeh et al., 2025; Wang and Fan, 2025). However, in the NLP field there is scarce scientific evidence on how effective and useful are LLMs to address this type of clinically (and other real-world) relevant cases.

Benchmark evaluations of LLMs such as Big Bench (Srivastava et al., 2023) or MMLU (Hendrycks et al., 2021) may not be representative of real-world, naturalistic scenarios and are not suitable for quantifying the efficacy of LLMs in clinical use cases; these benchmarks assess how well LLMs solve problems, or answer questions, across several domains; but they cannot assess the usefulness of LLMs to allow users to achieve their goals: If an LLM is used as a support counselor by adolescents suffering anxiety, no benchmark evaluation can assess if the LLM was effective in reducing the adolescents' anxiety symptoms.

In this paper, we provide a primer for biomedical NLP researchers on what is considered the *gold-standard* method for evaluating new treatments in clinical disciplines, namely Randomized Controlled Trials (RCTs) (Torgerson and Torgerson, 2008; Kamp et al., 2026; Lilliengren, 2023). RCTs, a type of clinical trial (Hackshaw, 2009), are the gold-standard due to their rigorous experimental

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 392–406
July 3-4, 2026 ©2026 Association for Computational Linguistics

design which allows researchers to control for biases and confounders that threaten to distort the true effect of the target treatment; thus, RCTs can lead to an unbiased and valid conclusion of the efficacy of a treatment (Bishop and Thompson, 2023; Hackshaw, 2009). We firmly believe that adopting RCTs will allow NLP researchers to stringently evaluate LLMs on real-world cases, and also to foster interdisciplinary research with clinical and healthcare communities. We truly hope this primer on RCTs will be widely used as a reference and guide in biomedical NLP in order to keep pushing the scientific progress of the field.

2 Background and Previous Works

2.1 Background

We start our account of RCTs by providing basic terminology. An RCT is an experiment, or study, conducted to provide a valid inference: Is a new treatment useful or effective for some goal when applied to participants? (Jadad and Enkin, 2007) A *treatment* is an intervention, a program, or in general, a mechanism which is hypothesized to provide a benefit to a group of people–the *participants* of the experiment. Similar to experiments in NLP where researchers test if adding a new module to an LLM helps the LLM to improve accuracy scores on a dataset, we seek to evaluate if our proposed treatment–for example an LLM–helps participants to attain a goal, such as distressed adolescents reducing their depression symptoms. In other words, we want to estimate the *effect* of the treatment, i.e. we want to quantify how much our treatment helped participants to attain a goal as measured by an outcome measure, such as a psychological questionnaire. This follows the same spirit as in NLP research where researchers estimate the effect of an improved or new model by comparing its mean accuracy score with that from a baseline model. However, when the evaluation of a treatment includes people, several *biases* and *confounders*–spurious factors–may be present which can distort the treatment effect threatening what is called the *internal validity* of the experiment, i.e. the capacity of a study to provide valid results. To counteract these spurious factors, researchers add another group of participants to the experiment called the *control group*–the equivalent of a baseline model in NLP–which will be compared to the *treatment group* (the group of participants on which we test our treatment), thus rendering valid results of the treatment's efficacy. In the following sections, we will provide the basic scientific principles of RCTs while showing how to avoid biases and confounders to design valid RCTs to evaluate LLMs on real-world scenarios.

2.2 RCTs for Evaluating Treatments in Clinical Disciplines

RCTs are extensively used across clinical disciplines. For example, in Psychiatry, RCTs have been used to evaluate treatments to improve the well-being of autistic children, such as fostering early social engagement (Carruthers et al., 2024), improving their socio-emotional and sleeping patterns (Boone et al., 2022), and developing their interaction and communication skills (Green et al., 2022). In Medicine, researchers have used RCTs to evaluate exercise programs for rehabilitation of people who have suffered a stroke (Moore et al., 2016), to assess interventions for promoting wound healing (Schneider et al., 2025), and to provide reliable evidence of the quality of new methods of cancer surgery (Ceelen and Soreide, 2023) and bone surgery (Jacobsen et al., 2020). Without RCTs assessing these type of critical treatments, end-users could be at risk of facing adverse consequences (Torgerson and Torgerson, 2008).

2.3 RCTs for Evaluating LLMs in Clinical Scenarios

Recently, RCTs have also been used to evaluate the efficacy of LLMs. For example, an LLM-based chatbot demonstrated comparable efficacy to a nurse hotline in reducing anxiety and depression symptoms of participants (Chen et al., 2025). Also, an RCT showed the usefulness of ChatGPT for physicians and medical students to conduct research but not any benefit for clinical decision-making (Weuthen et al., 2025). Furthermore, two RCTs showed that ChatGPT's role as a learning tool enhanced the performance of both dental and medical students for the lectures of clinical operative skills and orthopedics, respectively, compared to control groups (Huang et al., 2025; Gan et al., 2024). Remarkably, all of these evaluations come from medical venues. We strongly advocate for biomedical NLP researchers to be also actively engaged in these evaluations using this paper as a primer guide for learning the scientific principles for designing RCTs for LLMs.

3 Hypothetical Studies: Exemplifying the Evaluation of LLMs

To exemplify some points throughout the paper, we propose two hypothetical studies based on AI-enabled clinical support applications devised by the American Psychological Association (Association, 2025): 1) evaluating the efficacy of a tutor LLM for medical students aimed at improving their training in clinical diagnosis; 2) evaluating the efficacy of a complementary counseling LLM for mental health support aimed at reducing symptoms of distressed adolescents suffering from a psychological condition, such as depression or anxiety.

4 The Perils of Before-and-After Studies

It may seem straightforward to evaluate the efficacy of a new treatment by applying it on a group of participants (a treatment group) and measuring how much they benefited from the treatment based on an outcome measure. This study type is called a *before-and-after* or *pre-test-and-post-test* study (Ho et al., 2018; Sanders, 2019). Let us use our hypothetical example of a tutor LLM: We form one clinical diagnosis class in a school to apply the LLM as a treatment; we hypothesize that the LLM will cause this group to improve their test scores at the end of the study. Then, we assess our hypothesis by first obtaining baseline test scores–the *pre-test* scores–before the LLM is applied; then, we apply the LLM to the group to tutor them on diagnostic problems for a period of time; finally, at the end of the study we re-test the students to obtain their final scores–the *post-test* scores. One way to assess that this group benefited from the tutor LLM is by calculating how much their post-test scores increased with respect to the pre-test scores (Sanders, 2019): If the mean post-test score is higher than the mean pre-test score then we conclude that the LLM had a positive effect by helping the students to achieve such an improvement. However, this conclusion might be plagued with biases and confounders possibly rendering it invalid. These spurious factors could distort the test scores and we cannot disentangle their effects from the LLM's effect in this type of study (Moseley and Pinheiro, 2022). The solution to this problem is using an RCT. Nevertheless, a before-and-after study forms the basis of RCTs and we must learn which biases and confounders we must be cautious of.

4.1 Biases and Confounders

Biases are factors which systematically introduce an error in the form of a deviation to the treatment effect, but cannot be measured in specific units (Ho et al., 2018; Bishop and Thompson, 2023). On the other hand, confounders are measurable variables (unrelated to the treatment) which effect can be conflated with the effect of the treatment (Kingsley and Robertson, 2020) giving us the false impression that the treatment is (or not) useful. In before-and-after studies, confounders can usually be controlled–i.e. their effect can be neutralized; however, effects from biases can only be reduced–only RCTs can neutralize or eliminate them. Below we show the most relevant biases researchers must be aware of before any study is conducted.[1]

Placebo bias Participants volunteering for a new treatment mainly do so due to their willingness to recover from a problematic situation or condition. Evidence in the literature points to this willingness having a psychological effect on the outcomes of the participants regardless of receiving an effective or ineffective treatment; i.e. participants exhibit a placebo effect (Beins and McCarthy, 2018). This effect's magnitude could be so pronounced as to spuriously boost the treatment effect; for example, the adolescents in our hypothetical study may elicit a placebo effect after knowing that they will take part in an experimental group where the supporting counselor is a state-of-the-art LLM; this effect is elicited in response to their high expectation or enthusiasm in the LLM to help them alleviate their depression symptoms, effectively leading to a reduction in these symptoms, as it has been seen before in other cases (Bishop and Thompson, 2023; Torgerson and Torgerson, 2008). We note that the main mechanism of the placebo bias is still not fully understood (Piantadosi, 2017), but in our view, a possible explanation of this bias is that the participants mindset changes in a positive way, due to their high enthusiasm or motivation in receiving a novel treatment, aligning to the direction of the treatment effect; i.e. the participants hold the belief that the treatment will work, and they

[1] Other biases affecting the treatment that can be controlled via RCTs are the *Hawthorne* effect (Torgerson and Torgerson, 2008), the *maturation* bias (Marsden and Torgerson, 2012), the *demand characteristics* bias (Beins and McCarthy, 2018), the *regression to the mean* effect (Smith, 2015), the *history* bias (Marsden and Torgerson, 2012), the *experimenter* bias (Torgerson and Torgerson, 2008), or the *measurement* bias (Creswell and Guetterman, 2018). See Section A.1.

unconsciously exert a positive outcome by taking actions in parallel to the LLM's intervention such as changing daily-life habits or requesting extra-help outside the study, which will have a positive effect on the post-test scores regardless of whether the treatment really worked or not. This bias cannot be controlled in before-and-after studies. The optimal control is via an RCT (Section 6.2).

Test sensitivity or memorization bias Ideally, every participant in a before-and-after study takes two tests, a pre-test and a post-test, to assess their progress; however, this may induce a bias on the participants (Ho et al., 2018). For example, they may become sensitive to the pre-test in the sense of discovering what type of information or problems are used for evaluation; in this way, the participants may actively seek to learn any material aligned to that type of information or problems so as to intentionally improve the post-test evaluation giving the false impression that improvement in performance is due to the LLM when, in fact, it is due to the participants' sensitization and extra-effort (Marsden and Torgerson, 2012). Moreover, if the format or structure of the post-test is highly similar to that from the pre-test, then participants may get a better score just due to being familiarized with the test format or structure (Bishop and Thompson, 2023). This bias not only occurs in educational settings; it can also happen in clinical and psychological studies where participants may receive the same well-being survey or questionnaire as both pre-test and post-test–such as the Patient Health Questionnaire-9 (Yu et al., 2012) developed to assess depressive symptoms–in this case, participants may respond to the post-test questionnaire in a very similar way to how they responded to the pre-test questionnaire due to remembering their previous responses and unconsciously aligning to them (Navarro and Siegel, 2018). The best solution for this bias type is the RCT shown in Section 6.3.

Selection bias Another major threat to the internal validity of before-and-after studies is the problem of a biased sample of participants (Torgerson and Torgerson, 2008). The participants selected to the study may be biased towards a characteristic that could inflate or dilute the effect of the treatment. For example, if the recruitment is open to anyone wishing to take part in it, as is common in several studies, this will lead to one of the most common forms of selection bias: Recruiting people who are particularly enthusiastic to participate in a study mainly because they believe the new treatment can be a great help to their underlying problem or condition. Thus, this enthusiasm is a characteristic that has the potential to spuriously inflate the treatment effect via different mechanisms such as eliciting a placebo bias. In medical or psychological studies, other form of selection bias can take place: Participants willing to take part in a study may exhibit a degraded physical or psychological well-being which, if not profoundly serious, tends to naturally recover after given enough time regardless of receiving or not any treatment (Torgerson and Torgerson, 2008). These are just a few forms of selection bias; depending on the specific research problem many other forms can plunge into the study. RCTs are suited to control this bias.

Confounders effect Characteristics of the participants such as socioeconomic status, age, gender, etc. may have an impact on the post-test scores (Bishop and Thompson, 2023). For example, in our hypothetical study of a tutor LLM, recruiting only students from a high socioeconomic status risks confounding since access to high-cost curricular resources may influence post-test scores, distorting the LLM's effect. Similarly, in our hypothetical study of a counseling LLM, age effects can be confounded with the LLM's effect since younger (or older) participants may tend to recover more easily depending on the target condition. Hence, it is important to consider possible confounders, according to the specific problem, and control them, for example, via variation in the levels of the confounder, such as recruiting participants across varying ages, or via an RCT study.

5 RCTs: The Gold-Standard for Evaluations of LLMs

Since RCTs build on top of before-and-after studies, in this section we elaborate on their basic principles which allow for a strict control of the spurious factors occurring in before-and-after studies (shown in Section 4.1) and on the specific biases RCTs face as well as ways to control them.

5.1 RCTs Can Control Spurious Effects from Before-and-After Studies

The effects of biases and confounders in Section 4.1, which threaten the internal validity[2] of any before-and-after study, can be neutralized by

[2]See Section A.2 for a brief explanation of internal validity.

adding two key elements to this type of study: 1) a control group and 2) the randomized assignment of participants to control and treatment groups, where the treatment group is similar to the group in a before-and-after study; these elements constitute the basic and core structural features of a Randomized Controlled Trial (Creswell and Guetterman, 2018). In the following sections we elaborate on these characteristics.

5.2 Random Assignment of Participants to Groups: Controlling Biases and Confounders

The effects of biases and confounders can be neutralized, or counteracted, via the random allocation of participants to the control and treatment groups (Creswell and Guetterman, 2018). The goal with this action is to have the treatment and control groups nearly identical to each other in all participants characteristics and factors–including confounders and biases–except for the treatment to be evaluated; thus, the only difference between these two groups will be the effect of the treatment allowing us to easily estimate it. Simple randomization is a common method to achieve this as explained below.

Random assignment of participants to treatment and control groups is essential to balance the groups across observable and unobservable characteristics and factors that otherwise can lead to an invalid result. The simplest random method of allocation recommended in the literature is *tossing a coin* (Torgerson and Torgerson, 2008) where for each participant we obtain a computer-generated random binary label ("treatment" or "control"), or alternatively, we generate a random list of numbers, assign participants to numbers, and allocate odd numbers to one group and even numbers to the other. In any scenario, crucially, each participant must have the same chance to be assigned to any group. However, due to chance a confounder may not be fully balanced between control and treatment groups; to purposefully control this confounder we can resort to the method of *matched randomization* (see Section A.3).

5.3 RCTs Can Be Susceptible to Other Biases

RCTs may suffer from particular biases which are important to consider before conducting any study in order to propose any possible solution if required, as we show below (and in Section A.4).

Diffusion of treatment bias This bias, also known as *spillover effect* (Ariel et al., 2021), happens when participants in the treatment group share information of the treatment with participants in the control group in such a way that control participants can follow or adapt this information to their benefit having an effect on the post-test scores and thus threatening the study's validity (Creswell and Guetterman, 2018). A feasible policy to control this bias is to implement a double-blind setup, where neither the participant nor the researcher knows which participant is in which group (Bishop and Thompson, 2023; Friedman et al., 2015), in addition to a non-disclosure agreement where participants are requested to not sharing information from the study other than to relevant parties (such as the participants' parents if the participants are minors).

Resentful demoralization bias When participants assigned to the control group realize that they receive no treatment (or receive a traditional treatment which they believe is not effective), they may develop a feeling of resentment, demotivation, or demoralization which may have one or more consequences (Torgerson and Torgerson, 2008). One possible consequence is that control participants have a lower performance during the study than they usually do in normal days; this will have a negative effect on the post-test scores leading to artificially inflating the treatment effect. Another consequence is that control participants may seek for alternative treatments out of the RCT study; in our hypothetical example, students may seek help from a professor or engage with publicly available LLMs, such as ChatGPT, possibly leading to a spurious boost of their post-test scores. An additional consequence is that participants may quit the experiment due to a high impact of the demoralization or after rationalizing that their participation in the experiment will not lead to any benefit for them. A last possible consequence is that control participants may work harder during the study than they normally do due to their resentment of not being assigned any treatment (Creswell and Guetterman, 2018); this action will lead to inflated post-test scores which will minimize the true treatment effect. The resentful demoralization bias can be ameliorated or resolved either by a) giving a traditional treatment to the control group (such as a human tutor in our example) that can be regarded as an effective treatment (Beins and McCarthy, 2018), b)

reassuring control participants that they will have the same treatment as the treatment group after the experiment is finished (Bishop and Thompson, 2023), or c) replacing the control group with a placebo group (Torgerson and Torgerson, 2008) as we will elaborate on in Section 6.2.

6 Three Recommended RCT Designs to Evaluate LLMs

RCTs' experimental design can be adjusted to specific research questions, problems, and to analyze specific biases and confounders. We selected three main design types that we believe can be useful for biomedical NLP researchers to evaluate the efficacy of LLMs in real-world scenarios, from the basic *treatment vs. control* RCT design to more complex designs that aim to study the effect of relevant biases, namely the placebo bias and the test sensitivity or memorization bias which, we believe, can have a big impact on the outcome measure whenever studying the efficacy of LLMs across several domains, specially in educational and clinical contexts.

6.1 The Basic RCT Design: Treatment vs. Control Groups

A control group is most necessary to counteract the effects of several biases and confounders present in the treatment group via random allocation of the participants to these groups (Section 5.2). This design is the easiest and less costly to implement since only two groups of participants are required, where the suggested number of participants in each group is at least 50 in order to provide a precise estimate of the treatment effect (Navarro and Siegel, 2018). As mentioned before, only the treatment group receives the novel treatment and the control group receives either no-treatment or a traditional treatment (Beins and McCarthy, 2018). However, this design may not render the most precise treatment effect as the control group may not neutralize one possible bias, namely the placebo bias, since it tends to occur on participants who receive a novel treatment as it elicits the participants' high expectations that the treatment will be useful as imagined, overstating the treatment's effect (Piantadosi, 2017). This bias may or may not occur in an RCT; however, there is compelling evidence across fields, particularly Medicine and Psychology (Beins and McCarthy, 2018; Piantadosi, 2017), pointing to its frequent occurrence. To optimally control this bias it is suggested to add a placebo group.

6.2 Controlling the Placebo Bias: The Placebo Group

We introduced the placebo bias in Section 4.1. An optimal way to control this bias is to add a placebo group to the basic RCT design (or alternatively, to replace the control group with a placebo group) (Piantadosi, 2017). This group will receive a placebo[3] instead of a true treatment. The form of the placebo will depend on the form of the treatment under study; crucially, no placebo should provide any real treatment. For example, a placebo counterpart of a clinical diagnostics tutor LLM could be an LLM-based chatbot, in the style of the chatbot Eliza (Weizenbaum, 1966), that provides no support in diagnostics (and crucially, does not generate hallucinations), but responds to queries in a thoughtful-questioning way by asking the students what they think is a possible solution to a clinical problem, or why they think a candidate solution could be a good solution (or not).[4]

By adding the placebo group to our basic RCT design we now have the capability to not only remove the placebo effect from the treatment effect, but also to quantify it (Bishop and Thompson, 2023). More concretely, if we only desire to remove the placebo effect from the treatment effect, then we only need to subtract the mean post-test score of the placebo group from the mean post-test score of the treatment group since these two groups only differ in the treatment provided (novel treatment vs. placebo);[5] however, the placebo effect is mixed with all the other bias effects and we cannot quantify it. If we seek to quantify the placebo effect to develop an understanding of the strength of

[3] In simple and basic terms, a placebo is a mechanism mimicking or emulating a novel treatment, which also elicits high expectations from participants, but has no effect on them.

[4] We remark the potential complexity of proposing and implementing effective placebo counterparts of LLMs due to the inherent necessity of participants to interact with this placebo (e.g. via dialogues): It is difficult to mimic all the form or structure of the treatment without providing any functionality that could become in itself a treatment, while at the same time both not providing any negative effect and not allowing the participants to discover the status of the placebo (specially in long-term studies). For a deeper consideration of this issue and to help identifying an appropriate approach tailored to the needs of a specific study, we suggest previous works in the literature of Education (Hyldgaard, 2020; Adair et al., 1990; Aycock et al., 2018) and Psychology (Blease, 2018; Kirsch et al., 2016; Enck and Zipfel, 2019; Popp and Schneider, 2015; Gaab et al., 2019).

[5] This also removes the effects of all other biases and confounders explained in Section 4.1.

the participants' expectations regarding the efficacy of LLMs in real-world cases, then we simply subtract the mean post-test score of the control group from the mean post-test score of the placebo group–assuming that these two groups only differ in the placebo effect due to the optimal random allocation of participants to the three groups; nonetheless, for a more precise estimation, the recommended method is ANCOVA (Section 7.2).

6.3 The Solomon Design

This RCT design is considered as the ideal design to better quantify and understand the effect of the pre-test on the post-test scores (Solomon, 1949). This, in turn, allows us to tackle the test sensitivity or memorization bias (Section 4.1).

The Solomon design comprises 4 groups; it is a direct extension of the basic RCT study (Section 6.1) to which we add two more groups: Another treatment group receiving the exact same treatment, and another control group receiving either the same traditional treatment or non-treatment. However, these two groups will not receive a pre-test. Let us label our four groups where T_1 and T_2 are the two treatments groups, T_1 receiving a pre-test but not T_2, and C_1 and C_2 are the two control groups where C_1 receives a pre-test but not C_2. In this way, treatment groups only differ in the application of the pre-test, and similarly for control groups.[6] Then, we can estimate the effect of the pre-test on the post-test scores under two scenarios: When the treatment is present and when it is absent. Under the first scenario we compare the mean post-test scores of groups T_1 and T_2; if the mean score of T_1 is significantly larger than that of T_2 then we may ascribe this difference to the pre-test. Similarly, we compare the scores of C_1 and C_2; if there is no difference in this scenario, but there is a difference in the previous scenario, then we can claim that the pre-test affected the post-test not in a direct way but via the treatment (Ariel et al., 2021); for example, students became aware or sensitive of the type of information they must seek from the tutor LLM and thus exploited this sensitization to intentionally perform better on the post-test. Had there been a difference in scores between C_1 and C_2 in a similar magnitude and same direction as that between T_1 and T_2 then we could have claimed that the pre-test affected directly the post-test (possibly via a mechanism from the memorization bias, or due to test practice). This design requires twice as much participants as the basic design (around 200 is recommended (Navarro and Siegel, 2018)); but when the target population is abundant, this is the suggested RCT design to carry out.

[6] We note that assignment of participants to the four groups must be done using a random allocation method where any participant has the same chance to be assigned to any group.

7 Statistical Analysis of RCT Data

To estimate the effect of the treatment we can resort to two simple and popular statistical methods: The difference of means between treatment and control groups or an Analysis of Covariance (ANCOVA). We elaborate on these methods by exemplifying them with the basic RCT design (our account can be easily expanded to more groups).

7.1 The Simplest Method to Estimate the Treatment Effect: Difference of Means

In the basic RCT design, the only difference between treatment and control groups is the treatment effect (Murnane and Willett, 2010);[7] thus, by comparing the mean post-test scores of the two groups we estimate the magnitude of the treatment effect (similarly to how the mean accuracy score of a new NLP model is compared to that from a baseline model). For example, if students in our tutor LLM group obtained a mean post-test score of 8.4 points while the control group mean score was 6.5 points, then the treatment effect was 1.9 points. Alternatively, if we want to estimate the treatment's effect as the mean difference in progress from pre-test to post-test, then we can compare the average change of scores from pre-test to post-test between our two groups (Ariel et al., 2021); i.e., for each group we obtain the change in scores (post-test score − pre-test score) averaged across participants, and compare the two groups' means. (For other details see Section A.5.)

7.2 ANCOVA: Quantifying the Effects of the Treatment and Confounders

ANCOVA is a multivariate linear regression analysis where the dependent variable (post-test scores or change in scores from pre-test to post-test) is determined by a linear combination of independent variables, namely the treatment variable and covariates (i.e. confounders[8]) (Wang et al., 2019); co-

[7] Assuming both an optimal random allocation of participants to both groups and no placebo effect in the treatment group.

[8] Covariates is one name that confounders receive in statistics.

variates such as gender, age, socioeconomic status, etc. may have an effect on the post-test scores, specially when the sample size is small, since due to chance they may not be equally distributed across groups (Torgerson and Torgerson, 2008); moreover, including covariates in an ANCOVA model not only allows us to quantify their effect on the post-test scores, but also allows us to isolate them from the treatment effect (Read et al., 2013); furthermore, we can also estimate the effect of the pre-test on the post-test scores. This specification helps to improve the precision of the treatment effect estimation (Egbewale et al., 2014).

Equation 1 shows ANCOVA's model:

$$y = \mu + \beta_1(x_1) + \beta_2(x_2) + ... + \beta_T(x_T) + \epsilon \quad (1)$$

Variable y represents the dependent variable; term μ is the intercept of the linear model representing the expected score of a control participant when all covariates are set to zero. Coefficients β_i ($i \in \{1, 2, ..., n\}$) are parameters to be estimated and represent the effects of covariates x_i on the dependent variable; if a covariate is continuous, its coefficient represents the mean number of points that will be added to the dependent variable for each unit that the covariate increases; x_T is a binary indicator variable where $x_T = 1$ represents membership to the treatment group and $x_T = 0$ to the control group; thus, β_T is the estimated treatment effect which we interpret as the average number of points that a treatment participant will gain due to receiving the treatment. Interpretation of coefficients of categorical covariates is similar to that of the treatment variable. Finally, ϵ represents the error term.

Let us propose in Equation 2 an ANCOVA model obtained from our hypothetical RCT study testing the efficacy of our diagnostics tutor LLM where the dependent variable represents post-test scores.

$$score = 4 + 0.8(test) + 0.5(SES) + 2(LLM) \quad (2)$$

We include the covariate of socioeconomic status (SES) of students as an indicator variable where $SES = 0$ represents a low-SES student and $SES = 1$ a high-SES student; also, we model the effect of the pre-test score as a continuous variable; the treatment variable is a binary indicator where $LLM = 1$ indicates membership to the treatment group and $LLM = 0$ membership to the control group.

We interpret the model in Equation 2 as follows: The expected post-test score of a control student from low-SES with 0 points on the pre-test is 4 points. The effect of the treatment has a magnitude of 2 points, after adjusting for SES and pre-test; this means that students who have the support of the LLM ($LLM = 1$), compared to those who do not, score on average 2 points more on the post-test. Moreover, the pre-test has an effect on students possibly due to the sensitization or memorization bias depending on the exact form of the pre-test (Section 4.1); we interpret its coefficient as adding, on average, 0.8 points to the post-test score for each point obtained in the pre-test (after adjusting for SES and treatment group). Furthermore, after adjusting for pre-test and treatment group, high-SES students tend to obtain, on average, 0.5 points more on the post-test than low-SES students. (For details of ANCOVA see Section A.6.)

8 Ethical Concerns in RCTs for LLMs

LLMs are a recent technology. The study of their interaction with different user types across clinical domains has just started and significant further research is necessary to thoroughly understand the LLMs' effects; thereby, unexpected results may be obtained in any RCT study. Consequently, we recommend to first carry out RCTs on problems or conditions where the risks of adverse results are low, and avoid critical scenarios such as adolescents with severe depression since any negative side-effect could worsen their initial condition, such as becoming psychologically distressed or even suicidal. Critically, in the event where the LLM treatment were found to be adverse, it would be imperative to both assess the outcomes to decide whether to stop or continue the study and report these negative outcomes (Anderson et al., 2022) to allow the scientific community to ponder the adversities.

Moreover, RCTs are not without any ethical concerns. One common issue is the so-called *compensatory equalization* problem which refers to the unequal situation created under the basic RCT design (Section 6.1) or the placebo RCT design (Section 6.2), where only one group receives a potential benefit (a treatment), while the other group receives no-treatment and thereby no potential so-

lution to their underlying condition or problem (Creswell and Guetterman, 2018). This situation can be solved using either of the first two solutions proposed to control the *resentful demoralization* bias (Section 5.3), namely to give a traditional, helpful treatment to the control group (such as a human teacher or counselor if the novel treatment is an LLM), or to offer the novel treatment to the control (or placebo) group after the experiment has finished.

Another ethical concern in RCTs is the effect of the novel treatment. The usual hypothesis is that the novel treatment will provide a positive and helpful effect on the participants; however, in reality, the treatment effect may be negative or may have adverse secondary effects (Bishop and Thompson, 2023). Hence, it is recommended to continuously observe the participants to spot any sign of negative or adverse reactions in the participants' behaviors, and terminate the study if those reactions are found.

Finally, we strongly emphasize that any RCT study, prior to its deployment, must be reviewed and approved by an ethics committee from a university or an institutional board (Creswell and Guetterman, 2018), and consent must be obtained from all the parties involved in the study including the participants and any parents or legal guardians if required (Moher et al., 2010). Ultimately, participants have the right to drop out of the study at any time, and any action to force the participant to not do so must not be taken. We recommend (Moher et al., 2010) for a next guide on RCTs.

9 Conclusions

Evaluating the efficacy or usefulness of a new medical treatment, educational intervention, or therapy program in Medicine, Psychiatry, among other clinical fields, is done via Randomized Controlled Trials. This strict evaluation has received the nomenclature of evidence-based evaluation in those fields given that RCTs are the experimental method considered pertinent to provide rigorous scientific evidence for the efficacy of a target treatment. As exposed by Reiter (2025), the field of NLP is practically devoid of this type of evaluation for NLP applications. Crucially, we consider RCTs are necessary to assess recent NLP applications such as general-purpose LLMs–and more critically, biomedical or clinical support LLMs–which are no longer laboratory objects but everyday tools, and have shown both positive and negative impacts by helping users with routine tasks or by suggesting the ingestion of chemicals with potentially toxic characteristics, for example. As such, other fields, prominently Medicine, have recently started to evaluate LLMs across tasks and capabilities to provide scientific evidence of their benefit (or not) to specific populations. The NLP field is to take the next step in LLMs evaluations via RCTs as it is the designer and builder of these NLP tools. In this primer paper, we presented the basic principles and statistical methods required for designing Randomized Controlled Trials targeted to assess LLMs' efficacy for real-world cases, helping bio-NLP researchers to take the first step to conduct RCTs.

Limitations

As a primer paper, this work mainly shows content and scope limitations. It lacks advanced methods used in specific problems faced by RCTs such as better random assignment methods of participants to groups or a more comprehensive treatment of the placebo bias, more elaborate or complex RCT designs, recently discovered biases, a wider and more in-depth treatment of ethical concerns in RCTs, or state-of-the-art statistical methods for estimating treatment effects. Nonetheless, we note that advanced methods for each of these issues may be problem- or field-oriented; therefore, by providing a foundational treatment of the design of RCTs, this paper serves as the starting point to discovering advanced methods in specialized literature and to proposing methods tailored for the biomedical NLP field.

Ethical Considerations

We do not claim that this paper is a comprehensive, self-contained, and ultimate guide for designing and conducting RCTs. This is a primer paper that aims to introduce biomedical NLP researchers to the foundational principles of RCTs in order to advance to other specialized guides in other fields, to start the intellectual task of designing and proposing RCTs for evaluating LLMs, and to foster interdisciplinary works with researchers in clinical disciplines. Therefore, we advise the use of this primer paper as an introductory guide–but not the sole resource–for biomedical NLP researchers to start their study in designing RCTs for assessing LLMs while subsequently consulting more advanced texts and research papers.

Acknowledgments

We thank the reviewers for their valuable comments. This work was supported by Beijing Natural Science Foundation (IS25068).

References

John G. Adair, Donald Sharpe, and Cam Loi Huynh. 1990. The placebo control group: An analysis of its effectiveness in educational research. *The Journal of Experimental Education*, 59(1):67–86.

Alan Agresti and Maria Kateri. 2022. *Foundations of Statistics for Data Scientists: With R and Python*. CRC Press.

J. Michael Anderson, Conner Howard, Jessica Hardin, Cole R. Phelps, Chad Hanson, Reece M. Anderson, Matt Vassar, and Jake X. Checketts. 2022. Harms-related data are poorly reported among randomized controlled trials underpinning the american academy of orthopaedic surgeons clinical practice guideline recommendations for rotator cuff injuries. *Journal of Shoulder and Elbow Surgery*, 31(12).

Barak Ariel, Matthew Bland, and Alex Sutherland. 2021. *Experimental Designs*. SAGE Publications Ltd.

American Psychological Association. 2025. Artificial intelligence in mental health care. Last accessed: 2026-05-17. https://www.apa.org/practice/artificial-intelligence-mental-health-care.

Dawn M. Aycock, Matthew J. Hayat, Ashley Helvig, Sandra B. Dunbar, and Patricia C. Clark. 2018. Essential considerations in developing attention control groups in behavioral research. *Research in Nursing & Health*, 41(3):320–328.

Bernard C. Beins and Maureen A. McCarthy. 2018. *Research Methods and Statistics in Psychology*, 2 edition. Cambridge University Press.

Dorothy V. M. Bishop and Paul A. Thompson. 2023. *Evaluating What Works: An Intuitive Guide to Intervention Research for Practitioners*. Chapman and Hall/CRC.

Charlotte R. Blease. 2018. Psychotherapy and placebos: Manifesto for conceptual clarity. *Frontiers in Psychiatry*, 9.

Kelly M. Boone, Mark A. Klebanoff, Lynette K. Rogers, Joseph Rausch, Daniel L. Coury, and Sarah A. Keim. 2022. Effects of omega-3-6-9 fatty acid supplementation on behavior and sleep in preterm toddlers with autism symptomatology: Secondary analysis of a randomized clinical trial. *Early Human Development*, 169:105588.

Sophie Carruthers, Andrew Pickles, Tony Charman, Helen McConachie, Ann Le Couteur, Vicky Slonims, Patricia Howlin, Rachel Collum, Erica Salomone, Hannah Tobin, Isobel Gammer, Jessica Maxwell, Catherine Aldred, Jeremy Parr, Kathy Leadbitter, and Jonathan Green. 2024. Mediation of 6-year mid-childhood follow-up outcomes after pre-school social communication (pact) therapy for autistic children: randomised controlled trial. *Journal of Child Psychology and Psychiatry*, 65(2):233–244.

Wim Ceelen and Kjetil Soreide. 2023. Randomized controlled trials and alternative study designs in surgical oncology. *European Journal of Surgical Oncology*, 49(8):1331–1340.

Chen Chen, Kok Tai Lam, Ka Man Yip, Hung Kwan So, Terry Yat Sang Lum, Ian Chi Kei Wong, Jason C Yam, Celine Sze Ling Chui, and Patrick Ip. 2025. Comparison of an ai chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: Pilot randomized controlled trial. *JMIR Hum Factors*, 12:e65785.

John Creswell and Timothy Guetterman. 2018. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*, 6 edition. Pearson.

Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the New Statistics*, 1 edition. Routledge.

Bolaji E. Egbewale, Martyn Lewis, and Julius Sim. 2014. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Medical Research Methodology*, 14.

Audrey Eichenberger, Stephen Thielke, and Adam Van Buskirk. 2025. A case of bromism influenced by use of artificial intelligence. *Annals of Internal Medicine: Clinical Cases*, 4(8):e241260.

Paul Enck and Stephan Zipfel. 2019. Placebo effects in psychotherapy: A framework. *Frontiers in Psychiatry*, Volume 10 - 2019.

Xinyu Feng, Lidan Tian, Grace W K Ho, Janelle Yorke, and Vivian Hui. 2025. The effectiveness of ai chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: Systematic review and meta-analysis. *J Med Internet Res*, 27:e79850.

Lawrence M. Friedman, Curt D. Furberg, David L. DeMets, David M. Reboussin, and Christopher B. Granger. 2015. *Fundamentals of Clinical Trials*, 5 edition. Springer Cham.

Jens Gaab, Joe Kossowsky, Ulrike Ehlert, and Cosima Locher. 2019. Effects and components of placebos with a psychological treatment rationale – three randomized-controlled studies. *Scientific Reports*, 9(1):1421–1428.

Wenyi Gan, Jianfeng Ouyang, Hua Li, Zhaowen Xue, Yiming Zhang, Qiu Dong, Jiadong Huang, Xiaofei Zheng, and Yiyi Zhang. 2024. Integrating chatgpt in orthopedic education for medical undergraduates:

Randomized controlled trial. *Journal of Medical Internet Research*, 26:e57037.

AlMothana Gasaymeh, Asma'a Abu Qbeita, Reham AlMohtadi, and Mohammad Beirat. 2025. Exploring education students' use of chatgpt for academic and personal purposes: insights from a developing country context. *Frontiers in Education*, Volume 10 - 2025.

Jonathan Green, Kathy Leadbitter, Ceri Ellis, Lauren Taylor, Heather L Moore, Sophie Carruthers, Kirsty James, Carol Taylor, Matea Balabanovska, Sophie Langhorne, Catherine Aldred, Vicky Slonims, Victoria Grahame, Jeremy Parr, Neil Humphrey, Patricia Howlin, Helen McConachie, Ann Le Couteur, Tony Charman, and 2 others. 2022. Combined social communication therapy at home and in education for young autistic children in england (pact-g): a parallel, single-blind, randomised controlled trial. *The Lancet Psychiatry*, 9(4):307–320.

Allan Hackshaw. 2009. *A Concise Guide to Clinical Trials*, 1 edition. John Wiley & Sons, Ltd.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Anthony M H Ho, Rachel Phelan, Glenio B Mizubuti, John A C Murdoch, Sarah Wickett, Adrienne K Ho, Vidur Shyam, and Ian Gilron. 2018. Bias in before-after studies: Narrative overview for anesthesiologists. *Anesthesia and Analgesia*, 126(5):1755–1762.

Munier Hossain. 2021. *Making Sense of Medical Statistics: A Bite Sized Visual Guide*. Cambridge University Press.

Siyu Huang, Chang Wen, Xueying Bai, Sihong Li, Shuining Wang, Xiaoxuan Wang, and Dong Yang. 2025. Exploring the application capability of chatgpt as an instructor in skills education for dental medical students: Randomized controlled trial. *Journal of Medical Internet Research*, 27:e68538.

Kirsten Hyldgaard. 2020. The placebo effect in education? evidence based educational practice and the psychoanalytic concept of transference. *Journal of the International Society for Teacher Education*, 24(2):7–18.

Michael Kjær Jacobsen, Andreas Killerich Andresen, Annette Bennedsgaard Jespersen, Christian Støttrup, Leah Y. Carreon, Søren Overgaard, and Mikkel Ø Andersen. 2020. Randomized double blind clinical trial of abm/p-15 versus allograft in noninstrumented lumbar fusion surgery. *The Spine Journal*, 20(5):677–684.

Alejandro R. Jadad and Murray W. Enkin. 2007. *Randomized Controlled Trials: Questions, Answers, and Musings*, 2 edition. John Wiley & Sons, Ltd.

Caroline Barkholt Kamp, Sebastian Simonsen, and Sophie Juul. 2026. External validity of randomised clinical trials in psychiatry: borderline personality disorder as an example. *The British Journal of Psychiatry*, page 1–3.

Barbara Kingsley and Julia Robertson. 2020. *Your A to Z of Research Methods and Statistics in Psychology Made Simple*. Oxford University Press.

Irving Kirsch, Bruce Wampold, and John M. Kelley. 2016. Controlling for the placebo effect in psychotherapy: Noble quest or tilting at windmills? *Psychology of Consciousness: Theory, Research, and Practice*, 2(3):121–131.

Quan Li. 2018. *Using R for Data Analysis in Social Sciences: A Research Project-Oriented Approach*. Oxford University Press.

Peter Lilliengren. 2023. A comprehensive overview of randomized controlled trials of psychodynamic psychotherapies. *Psychoanalytic Psychotherapy*, 37(2):117–140.

Emma Marsden and Carole J. Torgerson. 2012. Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, 38(5):583–616.

S.E. Maxwell, H.D. Delaney, and K. Kelley. 2017. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 3 edition. Routledge.

Ryan K. McBain, Robert Bozick, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Joshua Breslau, Bradley D. Stein, Ateev Mehrotra, Lori Uscher Pines, Jonathan Cantor, and Hao Yu. 2025. Use of generative ai for mental health advice among us adolescents and young adults. *JAMA Network Open*, 8(11):e2542281–e2542281.

David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, P J Devereaux, Diana Elbourne, Matthias Egger, and Douglas G Altman. 2010. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340.

Sarah A. Moore, Djordje G. Jakovljevic, Gary A. Ford, Lynn Rochester, and Michael I. Trenell. 2016. Exercise induces peripheral muscle but not cardiac adaptations after stroke: A randomized controlled pilot trial. *Archives of Physical Medicine and Rehabilitation*, 97(4):596–603.

Anne M Moseley and Marina B Pinheiro. 2022. Research note: Evaluating risk of bias in randomised controlled trials. *Journal of Physiotherapy*, 68(2):148–150.

Richard Murnane and John Willett. 2010. *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

Mario Navarro and Jason Siegel. 2018. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, chapter Solomon Four-Group Design. SAGE Publications, Inc.

Steven Piantadosi. 2017. *Clinical Trials: A Methodologic Perspective*, 3 edition. Wiley.

L. Popp and S. Schneider. 2015. Attention placebo control in randomized controlled trials of psychosocial interventions: theory and practice. *Trials*, 16(1):150–152.

Kendra L. Read, Philip C. Kendall, Mathew M. Carper, and Joseph R. Rausch. 2013. Statistical methods for use in the analysis of randomized clinical trials utilizing a pretreatment, posttreatment, follow-up (ppf) paradigm. In *The Oxford Handbook of Research Strategies for Clinical Psychology*. Oxford University Press.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

S. Sanders. 2019. A brief guide to selecting and using pre-post assessments. Technical report, American Institutes for Research, The National Technical Assistance Center for the Education of Neglected or Delinquent Children and Youth, Washington, DC.

Ekaterina Schneider, Cristóbal Hernández, Robert Brock, Monika Eckstein, Guy Bodenmann, Markus Heinrichs, Ulrike Ehlert, Severin Läuchli, and Beate Ditzen. 2025. Intranasal oxytocin and physical intimacy for dermatological wound healing and neuroendocrine stress: A randomized clinical trial. *JAMA Psychiatry*, 83(2):118–127.

Gary Smith. 2015. *Essential Statistics, Regression, and Econometrics*, 2 edition. Academic Press.

Richard L. Solomon. 1949. An extension of control group design. *Psychological Bulletin*, 46(2):137–150.

PM Spieth, AS Kubasch, AI Penzlin, BM Illigens, K Barlinn, and T Siepmann. 2016. Randomized controlled trials – a matter of design. *Neuropsychiatric Disease and Treatment*, 12:1341–1349.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

David J. Torgerson and Carole J. Torgerson. 2008. *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*. Palgrave Macmillan London.

Rebecca Tutino, Elizabeth Schofield, Rebecca M. Saracino, Leah Walsh, Emma Straus, and Christian J. Nelson. 2024. A survey of statistical methods utilized for analysis of randomized controlled trials of behavioral interventions. *Palliative and Supportive Care*, 22(2):221–225.

Gerard J.P. Van Breukelen. 2006. Ancova versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59(9):920–925.

Andrew J. Vickers. 2005. Parametric versus nonparametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, 5(35).

Fei Wan. 2021. Statistical analysis of two arm randomized pre-post designs with one post-treatment measurement. *BMC Medical Research Methodology*, 21(150).

Bingkai Wang, Elizabeth L. Ogburn, and Michael Rosenblum. 2019. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics*, 75(4):1391–1400.

Jin Wang and Wenxiang Fan. 2025. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(621).

Herbert I. Weisberg. 2010. *Bias and Causation: Models and Judgment for Valid Comparisons*. Wiley.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Stephen G. West, Naihua Duan, Willo Pequegnat, Paul Gaist, Don C. Des Jarlais, David Holtgrave, José Szapocznik, Martin Fishbein, Bruce Rapkin, Michael Clatts, and Patricia Dolan Mullen. 2008. Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98(8):1359–1366. PMID: 18556609.

Felix A Weuthen, Nelly Otte, Hanif Krabbe, Thomas Kraus, and Julia Krabbe. 2025. Comparison of chatgpt and internet research for clinical research and decision-making in occupational medicine: Randomized controlled trial. *JMIR Form Res*, 9:e63857.

Xiaonan Yu, Wilson W.S. Tam, Paul T.K. Wong, Tai Hing Lam, and Sunita M. Stewart. 2012. The patient health questionnaire-9 for measuring depressive symptoms among the general population in hong kong. *Comprehensive Psychiatry*, 53(1):95–102.

A Appendix

A.1 Further Biases in Before-and-After Studies

Hawthorne effect Firstly discovered in early work of optimization studies, this bias takes effect on the participants behavior by increasing their effort, workload, or productivity due to their awareness of taking part in a research program and being observed by a researcher (Torgerson and Torgerson, 2008). Thus, part of the results obtained are due to the artificially increased effort of the participants. For example, students in our hypothetical study will work harder than usual when receiving support from the tutor LLM, regardless of their inner motivation, just due to the pressure imposed by the study's monitoring apparatus.

Maturation bias When a treatment is applied over a sustained period of time (usually in the range of months), natural biological or psychological changes can happen to the study participants which may have an effect on the post-test scores (Ho et al., 2018); these changes go under the umbrella terms of *maturation effects* or *temporal effects* (Torgerson and Torgerson, 2008). For example, in our hypothetical study of a counseling LLM, a maturation effect that can bias the effect of the counseling LLM is the natural self-healing of the study participants over time (Bishop and Thompson, 2023); thus, any improvement in their well-being post-test scores could be simply due to this temporal effect rather than due to the support of the LLM.

Demand characteristics bias This type of bias is elicited from the study participants after learning that they will take part in a novel treatment with the goal of improving their skills or solving their underlying problem or condition (Beins and McCarthy, 2018). Participants actively over-engage in the study by increasing their motivation or eagerness which, in turn, may improve their focus and their performance on the study's tasks; for example, students in our hypothetical study may pay more attention to the LLM's outputs than they usually pay to their professors at class, may put more effort and time for out-of-study learning activities, or may ask for extra-help from peers or professors, and so on. This participants' quest to make the treatment effectively work can aid the participants in improving their post-test scores biasing the effect from the LLM alone.

Regression to the mean Some effects occurring in a study may happen only due to chance; these effects are termed *regression to the mean* (Marsden and Torgerson, 2012). For example, in our hypothetical study of the tutor LLM, there is some likelihood that due to chance some of the students will obtain a low pre-test score, and these students will naturally increase their score at post-test time due to their normal intellectual or cognitive capability; consequently, this improvement will be wrongly attributed to the LLM's help; in other words, students who, due to chance, scored particularly high or low at the pre-test will tend to score closer to the group's mean at the post-test (Marsden and Torgerson, 2012; Torgerson and Torgerson, 2008) biasing the effect of the LLM.

History bias There are so-called history events occurring outside the laboratory where the treatment is applied which are not controllable by the researcher. This bias occurs when these events, happening in parallel to the study, affect either the status of the treatment or the participants engagement in the target tasks impacting on the results obtained (Marsden and Torgerson, 2012; Ho et al., 2018). For example, adverse weather leading to temporarily closing the laboratory can directly affect the participants engagement in the study possibly leading to a distortion of the post-test scores and also of the treatment effect.

Experimenter bias In this case, it is the researchers executing the study the ones biasing the participants outcomes (Torgerson and Torgerson, 2008); for example, if researchers firmly believe that the support counseling LLM will be useful for diminishing the symptoms of adolescents suffering anxiety, then these researchers may unconsciously work towards fulfilling their believe by modifying the well-being questionnaires or giving additional support to the participants not planned before the study with the aim of generating outcomes that favor the counseling LLM.

Measurement bias This bias can be present when taking measurements, scoring a test, or when using an incorrect instrument to elaborate the tests (Creswell and Guetterman, 2018), depending on the research problem. For example, if a test (pre-test or post-test) is not correctly designed to measure the intended skill, ability, behavior, or mental/physical state of the participants, then the test will result in an invalid outcome giving a biased result; follow-

ing our hypothetical example of a counseling LLM, if the well-being questionnaire contains ambiguous questions related to the symptoms of the participant, the given answers may not be valid and so may biased the treatment effect.

A.2 Internal Validity of RCTs

The golden goal in any scientific experiment is to obtain valid results (Weisberg, 2010). When we test a new treatment to see its efficacy, we aim to obtain an unbiased relationship between the treatment and the results (post-test scores); i.e. we seek to answer the question: Is the new treatment the real cause of the results obtained? For example, is a counseling LLM really helpful to alleviate the symptoms of a group of adolescents suffering from depression? When answering this question, we expect that the effect obtained on the participants is a true and direct cause from the LLM and not an effect from a bias or a confounder. Hence, to answer this type of question we need an experimental design embodying a necessary characteristic: Internal validity (Beins and McCarthy, 2018), which refers to the capacity of providing valid conclusions, i.e. to neutralize the effects of spurious factors, such as those shown in Section 4.1, that can distort the true treatment effect. Arguably, RCTs' experimental design is one of the study designs with the highest internal validity (West et al., 2008) which is the main reason why RCTs are considered the gold-standard for evaluating new treatments in the fields of Medicine (Torgerson and Torgerson, 2008) and Psychiatry (Spieth et al., 2016), among other fields; and we hope RCTs will become the gold-standard for evaluating LLMs in the NLP field as well.

A.3 Matched Randomization

Using simple randomization we risk that due to chance we may end up with unbalanced groups in one (or more) confounder(s), especially when the sample size is small (30 participants or less (Torgerson and Torgerson, 2008)), which can lead to invalid results. For example, in our tutor LLM study, we hypothesize that socioeconomic status has a significant impact on post-test scores since students in a high socioeconomic status may have access to extra learning resources, such as private lessons or specialized books. Hence, not thoroughly controlling this confounder may give us an inflated (or diluted) treatment effect if students with a high socioeconomic status are mostly allocated to the treatment (or control) group. The method of matched randomization consists of pairing participants with the same level of a confounder and randomly assigning each participant from each pair to treatment and control groups to balance them (Creswell and Guetterman, 2018); this can be extended to more confounders. For example, if we hypothesize that besides socioeconomic status, gender is another confounder, then we create pairs of students across combined levels of these two characteristics: (high, low) $\times$ (female, male); for instance, we match two male, high-socioeconomic-status students, then toss a coin and randomly assign them to treatment and control groups; and similarly for all combinations of levels. Due to its simplicity, matched randomization is a popular method; however, when the number of confounders increases, matching can become restrictive since it may be difficult to find an exact match for a given participant across all levels. In this case, it is recommended to increase the sample size to at least 100 participants and instead implement simple randomization(Torgerson and Torgerson, 2008). Alternatively, we can use a statistical method to control confounders that only requires simple random allocation, as shown in Section 7.2.

A.4 Attrition Bias

There are events that are unforeseeable by the researcher, such as participants dropping out of a study. The primary way how attrition invalidates a study is by inducing selection bias, regardless of participants dropping out from the treatment, control, or both groups, since it breaks the balance of other types of bias and confounders imposed by the random assignment of these participants to the groups (Torgerson and Torgerson, 2008). Attrition occurring in both groups does not guarantee to keep the balance between them since the participants dropping out of one group can have different characteristics from those dropping out of the other group. Moreover, when attrition occurs in one group it is not recommended to remove participants from the other group which have similar observable characteristics since they may not have similar unobservable characteristics. For example, let us suppose that the youngest adolescents in our hypothetical study dropped out of the treatment group because they believed that the counseling LLM was not helpful for reducing their anxiety symptoms; removing the youngest participants from the control group to balance the groups again risks inducing selection bias if dropouts from the treatment group

differ from the youngest control participants in the level of socioeconomic status, or in any other characteristic. To reduce the impact of attrition, it is suggested to implement the policy of Intention-to-Treat (ITT) (Bishop and Thompson, 2023), which advises researchers to keep track of dropouts, with their consent, to obtain their post-test scores since including them in the analysis of results will keep the balance between treatment and control groups imposed by the random allocation. For a deep treatise on ITT we recommend (Torgerson and Torgerson, 2008).

A.5 Details of Estimating the Treatment Effect Via Difference of Means

The estimated treatment effect obtained via the difference of means is free from the effects of bias and confounders; however, it may still have been affected by random variation, i.e. by chance, (Hackshaw, 2009) present in several factors. To account for this randomness we need a statistical test; in this case a common test is the t-test (Murnane and Willett, 2010), which we use to obtain a statistic called the t-statistic (the treatment effect normalized by the variability between the groups), which in turn we use to estimate the p-value which we could interpret as the likelihood that the treatment effect could have been obtained by chance (Hackshaw, 2009). The logic behind the estimation of a p-value is that we compare our t-statistic against a distribution (a t-distribution) of differences between the two group means built under the assumption that, in fact, there is no difference between the two means (i.e. the treatment effect is zero); if our p-value is less than an established threshold then it means that it is unlikely that we would observe such a difference in means under this distribution, so we can claim that this effect seems not to be due to chance (Hossain, 2021). Alternatively, we can estimate a *confidence interval* which is a range of values in which there is certain likelihood that the population treatment effect falls into; if zero is not inside this range, then we can claim that the treatment effect seems to not be due to chance as well (Cumming and Calin-Jageman, 2016). We suggest (Hackshaw, 2009; Hossain, 2021) for an elaboration of this statistical method and (Agresti and Kateri, 2022; Li, 2018) to estimate the t-statistic and its p-value with a statistical software as is commonly done in the literature.

While this is the easiest method to estimate the treatment effect, it has a drawback if our sample size of participants was small: It cannot measure the effects of confounders that may be lurking in the results (due to chance imbalance) which alter the treatment effect. So, unless we have a big sample size of participants and we carefully apply random allocation, and even so, the recommended statistical method for analyzing RCT data is ANCOVA (Van Breukelen, 2006; Vickers, 2005).

A.6 Details of ANCOVA

ANCOVA is one of the predominant statistical methods to analyze RCT data (Tutino et al., 2024) due to 1) its simplicity (Wan, 2021), 2) its flexibility to account for both types of variables (continuous and categorical) which allows us to estimate the treatment effect (a categorical variable) and the covariates' effects (categorical or continuous variables) (Maxwell et al., 2017), and 3) we can also estimate the effect of the pre-test on the post-test scores which results in a more precise estimate of the treatment effect (Egbewale et al., 2014).

The use of ANCOVA assumes some properties such as the linear relationship between dependent and independent variables, the dependent variable being normally distributed, or the variability of the groups being homogeneous; however, ANCOVA has been shown to be robust to departures from such assumptions (Wang et al., 2019).

To evaluate the fit of an ANCOVA model, we can use the mean-squared error: The averaged squared difference between predicted and true post-test scores, or the R^2 metric which estimates how much of the variability in the post-test scores is explained by the treatment and covariates, i.e. it measures the proportional reduction in the difference between predicted and real scores (Agresti and Kateri, 2022). For a deeper treatise on ANCOVA, its evaluation, and its extension to more groups, we suggest (Wang et al., 2019; Maxwell et al., 2017), and (Agresti and Kateri, 2022; Li, 2018) for implementing ANCOVA with a statistical software.

Citation-Aware Continual Pre-Training for Biomedical Language Models

Masaki Asada, Tomoki Tsujimura, Tatsuya Ishigaki,
Shusaku Egami, Ken Fukuda, Hiroya Takamura,

National Institute of Advanced Industrial Science and Technology (AIST)
Correspondence: masaki.asada@aist.go.jp

Abstract

The biomedical literature contains rich structured knowledge, including citation links that encode relationships between scientific studies, but such information is typically ignored in standard language model pre-training. We propose a citation-aware continual pre-training method for decoder-only language models that incorporates citation graph information from PubMed into next-token prediction by placing citation-linked abstract pairs within a shared context. We evaluate our method on multiple biomedical QA benchmarks using two model families. Results show that citation-aware continual pre-training achieves higher average accuracy than both the original base models and citation-unaware pre-training across biomedical tasks.

1 Introduction

The biomedical literature constitutes a large and information-rich corpus, with PubMed indexing tens of millions of abstracts (Liang et al., 2021). The citation links among these documents encode a human-curated graph of conceptual relationships. When a researcher cites a prior study, it typically reflects some form of domain relevance, methodological connection, or thematic continuity between the two works. However, this structural information is largely ignored in standard language model pre-training pipelines, which treat each document as an independent sequence drawn uniformly at random.

The encoder-only era demonstrated that citation structure can be a powerful supervision signal. LinkBERT (Yasunaga et al., 2022) placed linked documents into shared BERT (Devlin et al., 2019) contexts and introduced a document relation prediction objective, yielding substantial gains on QA and multi-hop reasoning benchmarks. Its biomedical variant, BioLinkBERT, achieved state-of-the-art results on BLURB (Gu et al., 2021), confirming that awareness of inter-document relationships translates directly into improved clinical language understanding.

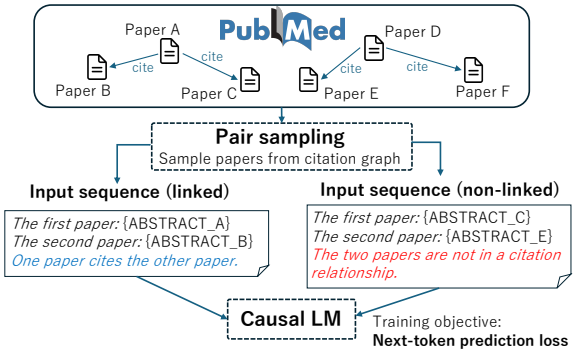


Figure 1: Overview of citation-aware biomedical pre-training. Each training example concatenates two abstracts with a relation text derived from the citation graph, and is trained with next-token prediction.

Decoder-only autoregressive large language models (LLMs), trained with a causal next-token prediction objective, have become the dominant modeling approach, and domain adaptation is typically pursued through continual pre-training on biomedical corpora such as PubMed Central. BioMistral (Labrak et al., 2024), for instance, adapts Mistral-7B with billions of biomedical tokens and reports improvements over general-purpose counterparts on medical QA benchmarks. Despite this progress, continual pre-training for causal LLMs remains document-independent, with each sample consisting of a single abstract or article and lacking structural signals about inter-document relationships. Citation-linked papers typically share concepts, methods, or disease entities, so placing them in a shared context window exposes the model to inter-document concept co-occurrences that no single abstract contains. Prior work has shown that citation structure encodes meaningful semantic relatedness at the document level (Cohan et al., 2020; Ostendorff et al., 2022); our hypothesis is that the same signal,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 407–412
July 3-4, 2026 ©2026 Association for Computational Linguistics

exposed at pre-training time through next-token prediction, supports integrative biomedical reasoning even when downstream questions are posed in closed-book form over a single passage. This raises a natural, to our knowledge unanswered question: does citation-aware pre-training confer the same benefits in causal LMs as it does for masked LMs?

As illustrated in Figure 1, we propose a method that incorporates information from the PubMed citation graph into next-token prediction pre-training. by constructing training samples that place citation-linked abstract pairs within the same context window together with a natural-language label indicating their relationship. We evaluate our approach on two small language models across multiple biomedical QA benchmarks, with a text-matched baseline that isolates the effect of citation structure from data quantity and lexical content. Our main contributions are as follows:

- We extend citation-aware pre-training from masked LMs to causal LMs by introducing a method that incorporates inter-document relationships while remaining compatible with standard next-token prediction training.
- We show that citation-aware continual pre-training improves accuracy over backbone models and citation-unaware pre-training across biomedical QA benchmarks and model families.

2 Related Work

2.1 Biomedical Pre-trained Language Models

Dedicated pre-training on biomedical corpora has a long history. BioBERT (Lee et al., 2020) first demonstrated that continued pre-training of BERT on PubMed yields substantial gains across named entity recognition, relation extraction, and QA. With the rise of decoder-only architectures, the community has adapted these insights to the autoregressive setting. BioGPT (Luo et al., 2022) is a GPT-2-scale causal LM pre-trained on PubMed abstracts, while PMC-LLaMA (Wu et al., 2024) scales this to 75 billion tokens from PubMed Central and medical textbooks. BioMistral (Labrak et al., 2024) demonstrates that continual pre-training of a strong general-purpose model on PubMed Central is competitive with purpose-built biomedical models at a fraction of the compute cost. More recently, MedGemma (Sellergren et al., 2025) extends the Gemma 3 architecture with medical-domain training on diverse clinical data including radiology images and medical text, while ELAINE-medLLM (Yano et al., 2025) adapts Llama-3-8B for the biomedical domain across English, Japanese, and Chinese through continual pre-training and supervised fine-tuning. Despite the diversity of these approaches, all treat the corpus as a flat collection of documents, ignoring citation relationships. Our work is the first to exploit PubMed's citation graph in causal LM continual pre-training.

2.2 Graph-Aware and Structured Pre-training

Citation links encode semantically meaningful relationships between documents, motivating a line of graph-aware pre-training methods. SPECTER (Cohan et al., 2020) learns document embeddings using citation links as a proxy for semantic similarity in a triplet-loss framework, and SciNCL (Ostendorff et al., 2022) refines this with neighborhood-contrastive learning over citation graphs. LinkBERT (Yasunaga et al., 2022) is most directly related to our work. It augments BERT's masked language modeling objective with a document relation prediction task, classifying whether two text segments are contiguous, randomly paired, or citation-linked. Applied to PubMed, BioLinkBERT achieves state-of-the-art performance on PubMedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2019) and the BLURB (Gu et al., 2021) benchmark. Our work translates the core insight of LinkBERT to the causal LM setting: rather than a classification auxiliary task, we encode document relation information as natural language tokens predicted autoregressively, making our method applicable to any decoder-only architecture without modification.

3 Method

3.1 Overview

We propose a citation-aware continual pre-training approach for decoder-only large language models. Our core idea is to exploit the PubMed citation graph, a large-scale human-curated signal of semantic relatedness between biomedical abstracts, by constructing training samples that place citation-linked document pairs within the same autoregressive context. Unlike LinkBERT (Yasunaga et al., 2022), which requires a separate classification head for document relation prediction, our formulation is architecturally transparent: the citation relation-

ship is expressed as a natural-language sentence appended to each paired-abstract sample and learned via standard next-token prediction.

3.2 Training Sample Construction

Given the PubMed citation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ corresponds to an abstract and each directed edge $(u, v) \in \mathcal{E}$ denotes that paper u cites paper v, we construct training samples of the following two forms.

Citation-linked sample:

```
The first paper: {ABSTRACT_A}
The second paper: {ABSTRACT_B}
One paper cites the other paper.
```

Non-linked sample:

```
The first paper: {ABSTRACT_C}
The second paper: {ABSTRACT_D}
The two papers are not in a
citation relationship.
```

For citation-linked pairs, we sample edges $(u, v) \in \mathcal{E}$ and place the abstracts of u and v as ABSTRACT_A and ABSTRACT_B, respectively. The order of the two abstracts within a pair is randomized to prevent the model from exploiting positional cues about citation directionality. For non-linked pairs, we sample two abstracts uniformly at random, verifying that no citation edge exists between them in either direction.

This formulation offers several practical advantages. First, it requires no modification to the model architecture or training objective: existing autoregressive pre-training pipelines can incorporate citation-linked samples without change. Second, the natural-language relation label is semantically informative; the model must associate the content of the two abstracts with the prediction of the concluding token sequence, implicitly encoding inter-document coherence.

3.3 Baseline: Citation-Unaware Pre-training

To isolate the effect of citation structure from the effect of data selection and lexical content, we construct a citation-unaware baseline as follows. Starting from the same set of paired-abstract samples used for our proposed model, we decompose each sample back into its constituent single-document units and shuffle these documents uniformly at random, forming standard single-document training samples. The resulting baseline corpus contains the same abstract texts as the proposed model but without citation-pairing information. The only difference between the two conditions is whether linked pairs are presented in context, enabling a clean ablation of citation structure.

4 Experiments

4.1 Experimental Settings

Backbone Models We apply our continual pre-training procedure to two recent small language models representing different model families.

- **Phi-3.5-mini-instruct** (4B parameters) (Abdin et al., 2024): a decoder-only language model pre-trained on a mixture of web and synthetic data.
- **Llama-3.2-3B-Instruct** (Dubey et al., 2024): an instruction-tuned decoder-only language model from the Llama 3 family.

Training Data Composition We set the total number of tokens after tokenization to approximately 1.8 billion, using the same set of PubMed abstracts for both citation-aware and citation-unaware settings to ensure that any performance differences cannot be attributed to differences in data quantity.

In the citation-aware setting, the total number of training tokens is 1,877,396,650, while in the citation-unaware setting it is 1,807,869,680. Although both settings are constructed from the same set of PubMed abstracts, the citation-aware data contains slightly more tokens because it includes additional textual prefixes such as "One paper cites the other paper." or "The two papers are not in a citation relationship."

While our overall scale is smaller than that of prior biomedical language models such as BioMistral (Labrak et al., 2024), which was trained on 3B tokens, it remains on the same order of magnitude.

Benchmarks We evaluate on five established biomedical QA benchmarks implemented via `lm-evaluation-harness` [1], reporting accuracy on all tasks: **PubMedQA** (Jin et al., 2019), **MedMCQA** (Pal et al., 2022), **MedQA-4opt** (Jin et al., 2020), and medical subsets of **MMLU** (Hendrycks et al., 2021) (*anatomy*, *clinical knowledge*, *college biology*, *college medicine*, *medical genetics*, and

[1] https://github.com/EleutherAI/lm-evaluation-harness

Citation-Aware		BioMistral-7B† NA	BioGemma-4B† NA	Phi-3.5-4B-Inst NA	Phi-3.5-4B-Inst ✗	Phi-3.5-4B-Inst ✓	Llama-3.2-3B-Inst NA	Llama-3.2-3B-Inst ✗	Llama-3.2-3B-Inst ✓
MedQA-4opt		44.4	**64.4**	55.9	54.8	55.5	55.9	53.5	55.4
MedMCQA		43.9	**55.7**	53.3	52.8	53.6	54.2	53.9	54.7
PubMedQA		37.6	73.4	74.8	74.8	74.4	75.2	75.4	**76.4**
MMLU	anatomy	49.6	59.3	62.2	61.5	63.0	62.2	60.7	**63.7**
	clinical KG	60.9	71.3	64.2	65.3	64.2	75.1	76.2	**75.5**
	college biology	56.9	70.8	72.2	71.5	72.9	84.0	81.3	**84.7**
	college medicine	55.5	**65.3**	57.2	59.5	57.8	64.2	**65.3**	**65.3**
	medical genetics	61.7	**83.0**	76.0	75.0	76.0	75.0	74.0	76.0
	prof. medicine	55.1	**76.8**	69.5	70.2	72.8	74.6	72.4	73.2
Average		51.7	68.9	65.0	65.1	65.6	68.9	68.1	**69.4**

Table 1: Results on biomedical QA benchmarks without task-specific fine-tuning. "NA" indicates that the original released model is evaluated without additional continual pre-training. "✗" denotes Citation-Unaware continual pre-training, while "✓" denotes Citation-Aware continual pre-training (our proposed method). Blue values indicate improvements over the corresponding base model, while red values indicate performance degradation. **Boldface** highlights the best score in each row. † indicates scores reported from the original papers.

Citation-Aware		Phi-3.5-4B-Inst NA	Phi-3.5-4B-Inst ✗	Phi-3.5-4B-Inst ✓
PubMedQA		76.6	77.4	**78.6**
MedMCQA		58.2	58.9	**59.1**
MedQA-4opt		57.8	57.3	**58.1**
MMLU	anatomy	59.3	59.3	**63.0**
	clinical_KG	74.3	74.3	**76.2**
	college_biology	**85.4**	81.3	84.0
	college_medicine	64.7	65.9	**67.6**
	medical_genetics	75.0	73.0	**76.0**
	pro_medicine	71.0	71.0	**71.7**
Average		69.2	68.7	**70.5**

Table 2: Results on biomedical QA benchmarks with task-specific fine-tuning.

professional medicine). Average accuracy (AVG) across all nine evaluation settings is reported as our primary aggregate metric.

Following prior work (Labrak et al., 2024), we adopt a 3-shot evaluation protocol without task-specific fine-tuning as our default setting. In addition, for the Phi-3.5-4B model, we also report results with supervised fine-tuning on the target tasks to assess the impact of downstream adaptation.

4.2 Results

Tables 1 and 2 present the results of continual pre-training under citation-unaware and citation-aware settings across different model families and evaluation conditions.

Overall, citation-aware pre-training consistently achieves the strongest performance across models and settings. In Table 1, citation-aware models outperform both the original released models (NA) and the citation-unaware baselines (✗) in terms of average accuracy. For instance, on Phi-3.5-4B, citation-aware pre-training improves the average score from 65.0 (NA) and 65.1 (✗) to 65.6, while on Llama-3.2-3B, it increases performance from 68.1 (✗) and 68.9 (NA) to 69.4. These gains are also reflected at the task level, with consistent improvements observed across multiple MMLU subsets and QA benchmarks.

Importantly, citation-unaware continual pre-training does not reliably improve performance over the base models and can even lead to slight degradation in several cases. This suggests that simply continuing pre-training on biomedical text without incorporating structural signals may not be sufficient, and can potentially harm previously acquired capabilities.

Table 2 further examines the effect of supervised fine-tuning (SFT) for Phi-3.5-4B-instruct model. The advantage of citation-aware pre-training not only persists but becomes more pronounced after fine-tuning, achieving the highest average accuracy (70.5) compared to both the base model (69.2) and the citation-unaware variant (68.7). Notably, citation-aware models achieve the best performance in the majority of individual tasks, indicating that the benefits of citation-aware representations transfer effectively to downstream adaptation.

While the magnitude in any single configuration is modest, the gain of citation-aware over citation-unaware pre-training has a consistent positive sign across two backbones, the majority of the nine tasks, and both the zero-shot and SFT regimes. Such cross-setting consistency is more readily explained by a systematic effect of the training signal than by configuration-specific noise.

5 Conclusion

We proposed a citation-aware continual pre-training method for decoder-only language models, incorporating citation graph information from PubMed into standard next-token prediction by placing citation-linked abstract pairs within a shared context. Experiments on biomedical benchmarks show that this approach improves performance over both the backbone models and citation-unaware pre-training across model families.

In future work, we plan to (i) compare citation-aware pre-training against a topically matched non-cited baseline to isolate the contribution of citation structure from topical co-occurrence, (ii) evaluate on multi-hop biomedical benchmarks to test the inter-document mechanism directly, and (iii) extend the approach to full-text articles and finer-grained citation contexts.

Limitations

This work has several limitations. First, our experiments use relatively small models (3B–4B parameters) and a limited pre-training scale (approximately 1.8B tokens), and the gains, while consistent, are modest in absolute terms; scaling to larger models and corpora may yield different outcomes. Second, our citation-unaware baseline draws abstract pairs uniformly at random, so citation-linked pairs differ from it in both citation status and topical similarity. The observed gains may therefore partly reflect co-presentation of topically related documents rather than citation structure per se, and a more controlled comparison would match non-cited pairs on topical similarity (e.g., via MeSH overlap or embedding similarity). Relatedly, all five benchmarks we use are closed-book single-passage QA and do not directly probe inter-document reasoning; multi-hop or cross-document biomedical benchmarks would test the proposed mechanism more directly. Finally, our method operates only on abstract-level data and cannot capture finer-grained relationships expressed at the sentence or passage level, which may limit its ability to fully exploit citation information.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhentao Liang, Jin Mao, Kun Lu, and Gang Li. 2021. Finding citations for pubmed: a large-scale comparison between five freely available bibliographic data sources. *Scientometrics*, 126(12):9519–9542.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553–568. Springer.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Ken Yano, Zheheng Luo, Jimin Huang, Qianqian Xie, Masaki Asada, Chenhan Yuan, Kailai Yang, Makoto Miwa, Sophia Ananiadou, and Jun'ichi Tsujii. 2025. ELAINE-medLLM: Lightweight English Japanese Chinese trilingual large language model for biomedical domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4670–4688, Abu Dhabi, UAE. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

TrackList: Tracing Back Query Linguistic Diversity for Head and Tail Medical Knowledge in Open Large Language Models

Ioana Buhnila[♪] **Aman Sinha**[♡♪♫] **Mathieu Constant**[♪]
[♪]ATILF, CNRS, Université de Lorraine, Nancy, France
[♡]IECL, University of Lorraine, Nancy, France
[♫]Institut Strauss, Strasbourg, France
Correspondence: firstname.lastname@univ-lorraine.fr

Abstract

While humans can easily produce various types of answers, such as definitions, examples or paraphrases, Large Language Models (LLMs) struggle to provide correct answers to medical questions that require diverse answer formats. In this paper, we introduce TrackList, a fine-grained linguistic and statistical analysis pipeline to investigate the impact of the pretraining data on LLMs answers to diverse linguistic queries. We also propose RefoMed-EN, a medical dataset consisting of 6,170 human-annotated medical terms alongside their corresponding definitions, denominations, exemplifications, explanations, or paraphrases. We investigated whether the high or low frequency of a concept (head or tail knowledge) impacts the language model's performance for answering medical questions. We evaluated the quality of the LLM's output using syntactic and semantic similarity metrics, statistical correlations and embeddings. Results showed that the LLM's answer quality for definition-type questions is the highest, while for the exemplification-type being the lowest. Additionally, we showed that for definition-type medical questions ("What is multiple sclerosis?"), LLMs are prone to paraphrase more for popular medical concepts, and less on more specialized medical knowledge.

1 Introduction

As Large Language Models are becoming widely used in many question-answering tasks and are applied to different fields, such as health and medicine, the need to explain and interpret the generated responses has become crucial. In this paper, we evaluated the most widespread real world use-case of LLMs, which is open question-answering, meaning the model gives a free text answer to a user question (Shailendra et al., 2024). LLMs are most efficient in question-answering tasks that cover popular concepts (head knowledge) (e.g. *cancer*, *asthma*) (Mallen et al., 2023; Li et al., 2024a)

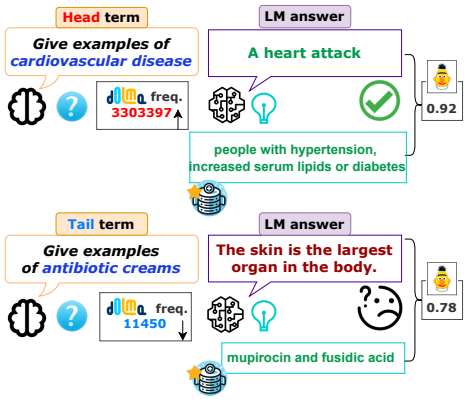


Figure 1: Language Models (LMs) tend to generate more hallucinated *definition-style* outputs even when directly asked to answer an *example-style* query for tail knowledge terms.

and when giving definitions to concepts. However, their performance drops when tested on torso or tail knowledge (*ideatory apraxia* or *erythematous angina*) (Sun et al., 2023; Kandpal et al., 2023) or when tackling complex questions (Daull et al., 2023). Our study analyzed the impact of query linguistic diversity on QA performance in the medical field, focusing on five different types of questions according to pragmatic functions of human communication (definitions, exemplifications, explanations, denominations and paraphrases). We demonstrated that LLMs perform worse when asked to give different types of answers, such as examples, to tail (or rare) medical knowledge (see Figure 1).

As black box and proprietary models such as ChatGPT showed limitations in terms of explainability and reproducibility (Zhao et al., 2024a; Liesenfeld et al., 2023; Ravichandran et al., 2024), more research should be conducted on open-source language models, to foster a deeper understanding of how LLMs process language and knowledge, especially in sensitive fields such as medicine. This study explored non-proprietary and open-source Large Language Models, OLMo (Groen-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 413–425
July 3-4, 2026 ©2026 Association for Computational Linguistics

eveld et al., 2024) and Pythia (Biderman et al., 2023). We chose these specific language models because they are fully open, thus fostering interpretability and reproducibility. These LLMs were released together with model weights, inference code, training/evaluation code, and pretraining corpus, DOLMA (Soldaini et al., 2024) for OLMo, and The Pile (Gao et al., 2020) for Pythia.

The research questions we investigated are the following: **(RQ1)** *Does the frequency of a medical concept in the pretraining data influence a language model's answer quality?* **(RQ2)** *How does the linguistic diversity and complexity of questions impact the language model ability to give accurate answers?* **(RQ3)** *To what extent are language models paraphrasing the data from their parametric memory for head and tail knowledge?* Based on previously research results on the general domain (Wang et al., 2024), we investigated whether frequent medical terms (head) lead to distributional memorization, and less frequent medical terms (tail) result in lower downstream task performance due to reliance on parametric knowledge.

The contributions of our paper are the following:

1. We proposed a fine-grained analysis and evaluation pipeline, **`TrackList`**, to **`tra`**ce ba**`ck`** query **`li`**nguistic diver**`s`**ity of head and **`t`**ail knowledge in Open Large Language Models. The pipeline is useful for evaluating LLMs' answers on various query types according to word frequency in the pretraining dataset. We will share the `TrackList` code with the NLP community to ensure reproducibility and to foster new research on LLMs' linguistic abilities for open QA tasks[1].

2. We share `RefoMed-EN`, an English dataset of 6,170 human annotated medical terms together with their context, discourse markers, corresponding paraphrases, lexical and pragmatic functions. The dataset is a non-contaminated benchmark extracted from human written texts and it is shared with open license together with the `TrackList` pipeline.

3. We conducted a detailed analysis of the LLMs' performance according to the linguistic complexity of the user question. We showed that LLMs have limited understanding of linguistic differences in queries, as they are more prone to giving definition-style answers to most queries, even when explicitly asked to output an example or a paraphrase for a medical concept.

[1]https://github.com/ATILF-UMR7118/TrackLIST

2 Related Work

Methods such as In-Context Learning (ICL) (Dong et al., 2024), Chain-of-Thought (CoT) prompting (Wu et al., 2023), and Retrieval Augmented Generation systems (RAG) (Lewis et al., 2020) have contributed to further interpret the LLM's generated response, and help mitigate hallucinations (Akbar et al., 2024; Huang et al., 2024; Asai et al., 2023). In the medical domain, LLM's output needs to be correct, therefore methods that combine ICL and RAG showed promising results, such as improving the accuracy of GPT4-Turbo's answers to oncology questions from 62.5 to 83.3% (with ICL) and 79.2% of questions (with RAG) (Iivanainen et al., 2024). Moreover, QA frameworks like Self-BioRAG that answer biomedical questions by reflecting and retrieving relevant documents showed a 7.2% improvement on average over the state-of-the-art open Small Language Models (SLM) of 7b or less (Jeong et al., 2024).

Large medical QA datasets like MedQuAD (Ben Abacha and Demner-Fushman, 2019) and MEDIQA (Abacha et al., 2019) gather different types of medical data, such as treatment, symptoms, definition, susceptibility or prevention. These datasets are built for medical experts, more than for the general public interested in understanding medical concepts (Nguyen et al., 2023). Moreover, these datasets are shared in a XML format, not allowing easy implementation for LLMs prompts. Furthermore, the dataset is divided in more medical than linguistic criteria. Therefore, our dataset, `RefoMed-EN`, is filling a gap in the benchmark leaderbord for query types in the medical field.

While evaluating LLMs answers to open questions is difficult, recent studies showed that human annotation is still needed when the generated answer does not match a gold standard (Kamalloo et al., 2023). In our work, we combine both human evaluation and automatic evaluation through similarity metrics that we further develop in Section 3.2. Determining what is head and tail knowledge is a research question in itself, as recent studies investigated this in detail (Mallen et al., 2023; Li et al., 2024b). We considered that the word frequency in the pretraining data as the popularity metric for our

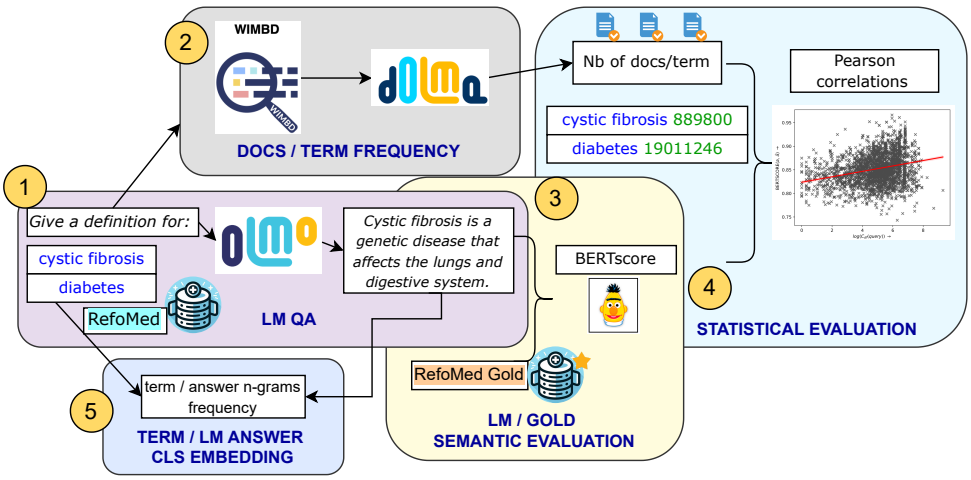


Figure 2: The pipeline of our method represented in five steps. 1) The zero-shot inference QA task using the medical concepts from the RefoMed-EN dataset. The dataset was divided in subdatasets according to the query type (detailed presentation in section 3.4). 2) We obtained the frequency in terms of number of documents for each RefoMed-EN concept. 3) We calculated the BERTScore between the LMs output and the RefoMed-EN gold standard. 4) These two values were used to compute Pearson correlations. 5) We computed a probability metric between the CLS embedding score between the term and the n-grams of the output, and their frequency in the pre-training corpora.

medical dataset.

Recent studies investigated whether LLMs memorize or generalize knowledge by exploring open language models. Wang et al. (2024) introduced the concept of distributional memorization, to measure the correlation between the LLM output probabilities and the frequency of the pretraining data. The authors evaluated the task performance with 3 to 5 n-grams search into The Pile (Gao et al., 2020), the pretraining corpus of the language model Pythia (Biderman et al., 2023). They found that LLMs generalize more in reasoning-intensive tasks, while they memorize more in simpler knowledge-intensive tasks (Wang et al., 2024). We test this hypothesis on the medical domain, to evaluate how knowledge specific concepts impact the LLMs answers' linguistic diversity.

Concurrent work investigated the concept of linguistic diversity in LLMs answers by evaluating their lexical, syntactic and semantic distribution compared to human linguistic richness (Guo et al., 2024). In our study, we investigated the linguistic complexity of LLM queries from a syntactic, semantic and pragmatic perspective, applied to a knowledge intensive task, medical QA. We present our method below.

3 Methodology

We illustrate our method in Figure 2. We conducted our experiments using four fully open language models, OLMo-1b, OLMo-7b, OLMo-7b-instruct, and Pythia-1b. We used the WIMBD tool's API (Elazar et al., 2023) to access the pre-training corpora DOLMA (for OLMo) and the `infini-gram` library (Liu et al., 2024) for The Pile dataset (for Pythia). We counted the number of documents that contain a certain medical concept, thus determining its frequency. We present our detailed rationale in the section below.

3.1 Formulation

We utilized a language model $\mathbb{M}_{\text{LLM}}$ and assumed access to a corpora C with $\mathcal{D}$ documents which was used to train $\mathbb{M}_{\text{LLM}}$. We represented the corpora C by a set of unique words $\{w_1, w_2, w_3, ...w_\infty\}$ present in at least one document $d \in \mathcal{D}$. Each of the document $d \in \mathcal{D}$ can be similarly represented by a subset of C as $\{dw_1, dw_2, dw_3, ...dw_\infty\}$ where dw_i represents word w_i present in document d. We defined a document-frequency count operator over corpora C and denote it by $\mathbf{C}_{df}$. This enabled us to count the number of documents present in corpora C which contained the provided term at least once[2].

We further utilized query-answer (Q, A) pairs where a question denoted by $q \in Q$ can be of different types as mentioned in Table 1 and an reference answer $a \in A$ which is borrowed from the RefoMed-EN dataset. In our experiment setup, we provides as input $q \in Q$ to the language model as

[2]For example, $\mathbf{C}_{df}$("*infection*" | C) outputs 543435456, which implies that there are 543435456 documents in C containing the term *infection* at least once.

follows:

$$\hat{a} = \mathbb{M}_{\text{LLM}}(q) \quad (1)$$

In the above equation, we marked the generated response from the $\mathbb{M}_{\text{LLM}}$ as $\hat{a}$. During evaluation, we considered different evaluation metrics (denoted by Ψ) to calculate the correctness (denoted by E) of the generated response $\hat{a}$ with respect to the gold reference $a \in A$, as follows:

$$E = \Psi(\hat{a}, a) \quad (2)$$

Lower value of E, for example, in case of BERTscore, implies less relevance of generated answer and vice versa.

3.2 Evaluation Metrics

As the goal of our study is to advance the interpretability of open large language models by describing their behavior, we chose simple and easy to interpret metrics that allowed us to analyze the relationship between query term frequency, gold answers, and generated LLM answers.

Tracing back semantic similarity. We calculated the BERTscore (Zhang et al., 2019) between the generated answers and the gold standard from the RefoMed-EN dataset.

Statistical correlation. We computed the Pearson correlation metrics to compare the BERTscore calculated before with the term's frequency in the pretraining corpus.

CLS cosine similarity. We computed the cosine similarity of sentence embeddings to compare the semantic similarity between the medical term and the n-grams of the generated answer.

3.3 Linguistically Annotated Dataset

We constructed RefoMed-EN by automatically translating RefoMed (Buhnila et al., 2024) from its original version in French to English using a licensed DeepL Translator API. RefoMed is a French annotated dataset of 6,170 annotated medical terms with their corresponding reformulations. The dataset is comprised of paraphrases and reformulations semi-automatically extracted from scientific and popularization medical texts and abstracts from the ClassYN (Todirascu et al., 2012) and CLEAR (Grabar and Cardon, 2018) French medical corpora. The dataset required extensive pre-processing as there were formatting and translation inaccuracies from the original annotation. We chose to translate this dataset because there is no freely available benchmark annotated on question types following pragmatic and linguistic theories for the medical field (to the best of our knowledge). More information about the different types of linguistic annotations can be found in Appendix A. The translation of a French dataset assures that there is no benchmark contamination between the pretraining corpora and the test dataset (Sainz et al., 2023; Li et al., 2024c). We share this new non-contaminated benchmark with the NLP community on GitHub. We present below the question types explored in this study.

3.4 Query Linguistic Diversity

Linguistically, the concept of *reformulation* is defined as textual and discursive act performed with a precise objective (Grabar and Eshkol, 2016). *Reformulations* have a well defined pragmatic role by expressing a content in a different semantic or lexical representation, adapted to a specific audience and communicational need, like science popularization or education. The linguistic diversity is correlated to the pragmatic usage of language in humans, such as asking for a definition, an explanation, reformulation or paraphrase of a concept, or to receive examples of a certain concept (Grabar and Eshkol, 2016).

In this work, we analyzed the role of reformulations in the case of medical knowledge popularization for laypeople and patients (Grabar and Eshkol, 2016). We focused on a knowledge intensive question-answering task taking into account the five most common types of questions that require reformulation processes, as shown with examples in Table 1. We present the number of questions by type from the RefoMed-EN dataset in Table 2.

4 Experimental Setup

This section presents the experiments we conducted to evaluate the generated text according to query types and term frequency. Experiments were done on a P100 NVIDIA GPU, for an individual runtime of $\leq$ 15 hours including different steps involved.

4.1 Linguistic Diversity and Frequency for Task Performance

We explored the impact the query linguistic diversity has on the quality of the generated answer. We evaluated the task performance of OLMo and Pythia models in a zero-shot QA setting by conducting several experiments:

Pragmatic function	Definition	Query type	Example
`Definition` (**DEF**)	a difficult or technical concept is defined to ease comprehension	Give a definition for / What is *multiple sclerosis*	Multiple sclerosis (MS) is a *disease of the nervous system of immune origin*
`Exemplification` (**EX**)	the meaning of a concept is illustrated through examples of types and subtypes of entities	Give examples of *heart and vascular diseases*	Heart and vascular diseases include *heart attacks, angina, stroke, sudden cardiovascular death and the need for heart surgery*
`Denomination` (**DEN**)	a concept is reformulated through a semantically similar concept, without simplification	Give another denomination for *motor neuron disease*	Pharmacotherapy for pain management in *amyotrophic lateral sclerosis* (motor neuron disease)
`Paraphrase` (**PARA**)	the concept is reformulated through a easy to understand semantically similar synonym	Give a paraphrase for *hypotension*	Hypotension, i.e. *low blood pressure*, frequently occurs in newborns
`Explanation` (**EXP**)	the concept is explained through its process or a part of it	Give an explanation for *autoimmune diseases*	These are autoimmune diseases that can be explained by the fact that *the organism produces an antibody against the person's skin*

Table 1: Query types, definitions and examples according to pragmatic functions annotated on the RefoMed-EN corpus. These queries were used for prompting in our QA task.

	DEF	EX	DEN	PARA	EXP
`ClassYN EX`	207	305	131	30	32
`CLEAR EX`	919	379	401	73	67
RefM-EN-EX	*1126*	*684*	*532*	*103*	*99*
`ClassYN GP`	862	343	235	285	149
`CLEAR GP`	883	470	175	124	100
RefM-EN-GP	*1745*	*813*	*410*	*409*	*249*
RefM-EN	**2871**	**1497**	**942**	**512**	**348**

Table 2: Distribution of question types across subcorpora in RefoMed-EN (RefM-EN). EX denotes expert-oriented texts; GP targets the general public. Question types are ordered by frequency (left to right).

- We prompted the LLM to give a short answer (**a**) to the given question (**q**) as an expert in the field: `"You are a medical expert. Answer the following question in a short sentence"`. The queries were divided by question type, as shown in Table 2.
- We evaluated the quality of the generated answers for each type of query by computing the BERTscore between the generated answer and the gold standard, the human annotated paraphrases and definitions from RefoMed-EN.
- We calculated the Pearson correlation between the BERTscore and the frequency of the medical concept in the pretraining corpora.

4.2 CLS Embbedings and Co-occurrence Probability

We investigated the link between words embeddings and frequency in the pre-training corpus by conducting the following experiments:

- Firstly, we computed the **cosine similarity score** between the embeddings of the simple or multi-word medical term to be explained or defined, and the embeddings of the answer generated by the language models. We used the sentence-transformer model `paraphrase-MiniLM-L6-v2`. In order to do an exhaustive comparison, we dissolve the reference answer (a) and generated answer ($\hat{a}$) into all possible n-grams. Then, we compute the pairwise cosine similarity between 2, 3, 4 and 5 possible n-gram pairs.
- Secondly, we calculated the **probability score** between two document-frequency counts: 1) the frequency of the query term (q) in the corpora (C), and 2) the frequency of the term together with the generated answer of the language model in the corpus. We used the `WIMBD` tool to found the frequencies in terms of number of documents from DOLMA, and `infini-gram` library for The Pile.

$$P_{cooccurence} = \frac{\mathbf{C}_{df}(q|\hat{a}, C)}{\mathbf{C}_{df}(q|C)} \quad (3)$$

We calculated different n-gram combinations, of 2 to 5 n-gram length. We kept the top-3 semantically meaningful values. We show the statistical correlation between these variables in the Results section.

4.3 Tracing Back Head and Tail Knowledge

We listed the most popular medical concepts (`head`) as those appearing most frequently in the parametric memory, and the least popular medical concepts (`tail`) as having the smallest number of corresponding documents in the parametric memory.

Term	Docfreq	Examples
disease	75724477	A disease is a medical condition that affects the body's structure or function
cancer	50918098	A disease that affects the cells
anxiety	35107524	Anxiety is a feeling of fear and uneasiness
testosterone-inhibiting	115	The answer is testosterone-inhibiting
biological tissue damage	503	Give examples of biological tissue damage
lifelong neurological consequences	29	Explain the neurological consequences of the disease

Table 3: Examples of LLM paraphrasing. Frequent terms are paraphrased, while rare terms yield outputs similar to or derived from pre-training data.

We traced back 100 concepts (1.62%), where 50 were head concepts (0.8%) and 50 percent tail concepts, to the pretraining corpus. We discarded the very long tail concepts from RefoMed-EN that had zero frequency. However, it is important to note that WIMBD's search is exact match based (while ignoring special characters and punctuation), and that some zero frequency terms are very long and technical, such as "disorder of bronchial ventilation" or "SMN1 gene-related proximal spinal muscular atrophy". To compare expert (EX) to general public (GP) datasets, we split this number evenly between RefoMed-EN-EX and RefoMed-EN-GP. We analyzed word level frequencies in 100 documents for each term downloaded with WIMBD.

As the linguistic diversity and the quality of the pretraining data is extremely important in the task performance evaluation, we conducted a close up analysis of corresponding documents in the DOLMA corpus. We analyzed the texts for a selected number of `head` and `tail` concepts.

	DEF	EX	DEN	PARA	EXP
O1b	0.27	0.05	0.15	0.11	0.12
O7b	0.23	0.16	0.15	0.14	0.14

Table 4: Pearson Correlation coefficient (ρ<0.05) between BERTscore(a, $\hat{a}$) and $\texttt{log}(\mathbf{C}_{df}(q))$ for OLMo-1b (O1b) and OLMo-7b (O7b).

5 Results and Analysis

Statistical correlations between medical concepts frequency in the parametric memory. To answer **(RQ1)**, we computed the correlation coefficient between BERTscore and query term frequency C_{df}. We show the Pearson correlations scores coefficient between the BERTscore (a, $\hat{a}$) and log(C_{df} (q)) on the full dataset of 6,170 terms and gold paraphrases, in Table 4. The best scores were obtained with OLMo-1b on the DEF query type (0.27). While the value indicates a weak correlation, it is the highest among all query types, showing some link between the semantic similarity of the generated answer compared to the gold standard answer, and the frequency of the term in DOLMA. Second best scores are on DEN (0.15), while PARA and EXP have lower values (0.11 and 0.12). The lowest correlations score were obtained on the EX query type (0.05). We obtained similar results for the bigger model, OLMo-7b, for DEF (0.23), DEN (0.15), and EXP (0.14). OLMo-7b has better scores for EX (0.16) and PARA (0.14), which shows that the model handles the complex linguistic task better. We conducted a manual analysis of the quality of the generated answers for each query type, described in the next section that tackles **(RQ2)**.

When computed separately by type of corpus, expert medical texts (RefoMed-EN-EX) and general public medical texts (RefoMed-EN-GP), the results are consistent: the scores are the highest for DEF and DEN. However, the scores on RefoMed-EN-EX are better than those on RefoMed-EN-GP for DEF, and even higher than the values on the full dataset (0.31 compared to 0.24 and 0.27). This result might indicate that in the expert texts, the task performance of the model on DEF is slightly more correlated to the term frequency than in the general public (as RefoMed-EN-EX has a higher number of technical terms). On the contrary, for DEN, scores are higher on RefoMed-EN-GP (0.19) than RefoMed-EN-EX (0.11), also surpassing the full dataset score (0.15). We analyzed these results manually in the next section.

QA task performance evaluation on linguistically diverse query types. To investigate **(RQ2)**, we conducted a qualitative analysis of the LLM's linguistic *understanding*. Our hypothesis was that language models show very high performance for the `definition` type query, as this type of query

is very frequent in knowledge intensive QA benchmarks (Rebboud et al., 2024; Zhao et al., 2024b; Fei et al., 2023). The results are consistent with our hypothesis (Table 4). On the flip side, we observed that the LLM does not completely *understand* the concept of `paraphrase` or `explanation`, as it generated definitions as answers instead.

We noticed that OLMo-1b obtained lower task performance with the *denomination* and *exemplification* types of queries, as it tends to either repeat the term (for denomination in particular) or give a definition style answer for both. The difficulties for these two queries come from linguistic and domain specific linguistic traits, such as:

- `Denomination query` - the LLM repeats the term instead of giving another synonym for the medical concept. Furthermore, the presence of highly technical long-tail terms (such as *lymphocytic bacterial meningitis*; freq=2) or opaque abbreviations (e.g. *FixM/F*; freq=11) renders the task even more difficult for the language model.

- `Exemplification query` - the LLM does not output examples (i.e instances, types, subtypes) when the query term is very technical, thus long-tail knowledge: (*Give examples of severe motor disorders* → *The answer is severe motor disorders*). However, the model task performance increases for head knowledge (*1. Give examples of cardiovascular disease* → *a heart attack*; *2. Give examples of psychological factors* → *The answer is: Psychological factors include: fear, anxiety, and depression*).

Tracing back head and tail knowledge in the pretraining corpus. We show the 100 analyzed terms and their document and term frequencies in Table 6 and Table 7 in Appendix B. We looked into a total of 5,074 documents from the DOLMA corpus, where 5,000 contained the `head` words and 74 the `tail` for the tail concepts. For each `head` term out of the 50, we downloaded 100 documents where this term appeared at least once. For `tail` terms, we looked for concepts that appeared in 1 to 3 documents in the DOLMA pretraining corpus.

The `long tail` terms with very low frequency (>4) are very technical long multi-word terms, such as "pseudo-rhizomelic polyarthritis", "lymphocytic bacterial meningitis", "CoA HMG reductase inhibitors". However, there are also very long multi-word terms in RefoMed-EN-GP, like "work-related musculoskeletal disorders of the upper limbs and neck" or "shiftworker sleep disorder" with a term frequency of 2 and 1 in the whole DOLMA corpus.

Criteria	Model	RefM-EN-EX	RefM-EN-GP	RefM-EN
Diversity	OLMo-1b	-0.3904	-0.5710	-0.4403
	Pythia-1b	-0.0206	-0.1202	-0.0317
Scalability	OLMo-7b	-0.3656	-0.3308	-0.3501

Table 5: Pearson correlation between CLS cosine similarity (query term, response-top3-ngrams) and the cooccurence probability of frequency of the term together with its top3 n-grams. (Red color denotes ρ>0.05)

Regarding the `head` terms, the most frequent term from RefoMed-EN is "life", which appears in 497M documents in the pretraining corpus, while the second most frequent is "control" with 230M documents. Other head terms are "conditions" (130M), "function" (114M), "skills" (126M) and typical medical head concepts such as "treatment" (113M), "pain" (80M), "disease" (75M) and "cancer" (50M). While analyzing the term frequency, we noticed that terms such as "control", "function" and "client" appear in code texts, thus irrelevant for your medical knowledge QA task.

CLS n-gram embeddings on head and tail terms. We explored **(RQ3)** in line with Wang et al.'s (2024) hypothesis: LLMs generalize more in reasoning-intensive tasks, and they memorize more in simpler knowledge-intensive task (like in the case of tail knowledge). We conducted the CLS n-gram analysis on the list of 100 head and tail concepts to verify this hypothesis in the medical field. We calculated the CLS embedding BERTscore between the medical term and n-grams of the generated answer of different sizes (2, 3, 4 and 5 n-grams). We kept the top 3 best BERTscore for each term and we compared them with the frequencies of the term together with theses n-grams in the pretraining corpus. We show the distribution of the results in Table 5.

The negative distribution scores (-0.44 on RefoMed-EN for OLMo-1b) indicate that the language model tends to create its own sentences and does not take full information package directly from DOLMA. In terms of scalability, OLMo-7b is inline with OLMo-1b (-0.35 on RefoMed-EN), while another family of models, Pythia-1b, show very low distribution scores. This suggests that the LLM prioritizes semantic paraphrasing (backed by the semantic evidence in relation to the CLS similarity) (see examples in Table 3), and it is less likely to reproduce the same content from the pretraining corpus, as proven by the syntactic evidence related to the n-gram occurrence. OLMo-1b's scores are

better for head terms and for the medical texts destined to general public texts (-0.71). This finding demonstrated that the model used more diverse vocabulary for popular medical knowledge, and less technical medical concepts, particularly in the expert texts (+0.07). For the bigger model, OLMo-7b, we obtained similar values for both types of texts, popular and scientific. This observation supports previous studies on general language (Wang et al., 2024), indicating that the LLM memorizes more on knowledge intensive questions in the medical domain, than for popular medical terms.

6 Discussion

Scaling up small language models does not necessarily improve performance. We replicated our experiments on a bigger model, OLMo-7b. We notice a similar trend for definition-type questions: both OLMo-1b and OLMo-7b exhibit positive Pearson correlation coefficient (p) between BERTscore and query term frequency. Interestingly, we noticed that OLMo-7b shows a weaker trend (p=0.23) compared to OLMo-1b model (p=0.27), this implies that with increase in the size of language models the ability to deal with tail knowledge does not necessarily increase. Regarding overall comparison of CLS n-grams scores on head and tails (see Table 5), we further notice a similar trend between OLMo-1b and OLMo-7b models. However, these findings might change when testing even bigger models.

Human evaluation shows hallucinations are not scale related. We conducted a manual analysis of 400 answers given by the four language models for the 100 head and tail terms dataset (Figure 3). We evaluated task performance and diversity of models by looking directly into the data for hallucinations. Our analysis showed that Pythia-1b is more prone to generating hallucinatory texts (+22%) compared to OLMo-1b, as it was previously shown (Groeneveld et al., 2024). As for bigger models, OLMo-7b is hallucinating more (+19%) than its Instruct version. The best model remains OLMo-1b (33% hallucinated answers), followed by OLMo-7b-Instruct (39%). Pythia-1b demonstrates an opposite trend for definition-type questions (DEF) as compared to OLMo-1b, with a comparatively low Pearson correlation coefficient of -0.103. However, OLMo-7b shows a similar positive trend for paraphrase-type questions (PARA) with a coefficient of 0.14.

We want to stress that results might differ with bigger models. Nevertheless, the main goal of our study was to analyze smaller sized LLMs to support reproducibility and frugality for responsible use of generative AI tools.

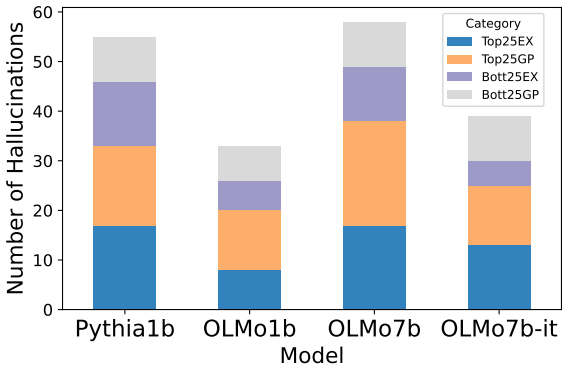


Figure 3: Manual analysis of hallucinations on 100 head and tail terms. Best model is OLMo-1b, with the lowest rate of hallucinations (33%).

More linguistic diversity metrics are needed. BERTscore has limitations in capturing the semantic aspects of query types. For example, for denomination-types queries where the LLM only repeated the medical term in the query, the BERTscore will be very high, but not relevant. In a concurrent study, Guo et al. (2024) compared lexical, syntactic and semantic distribution of LLMs texts to human gold answers. Lee and Lee (2023) proposed a pipeline to identify 220 popular handcrafted linguistic features. However, these features are not thoroughly tested on different writing tasks, and more research needs to be done on their utility for diverse query types.

7 Conclusion

Our study introduced TrackList, a pipeline to analyze, trace back and evaluate a language model's answer to diverse linguistic queries. We showed that frequency of terms in the pretraining corpus impacts performance: LLMs tend to give inaccurate answers for head terms and more accurate answers for tail terms in the medical domain. Furthermore, the LLMs studied were prone to paraphrase more on popular medical terms, and less on very scientific terms. Our linguistic analysis demonstrated that language models tend to give definition-type answers to different queries, even when asked to give examples or paraphrases. This demonstrates that Small Language Models have a limited linguistic *understanding* (or representation) of the best fitted answer for diverse question types.

Our study focused on small size language models to investigate the inner working of LLMs without domain knowledge finetuning, DPO finetuning or RAG systems. Our purpose was to analyze how the pretraining corpus and the word frequency (head and tail) impacts the accuracy of the small LLMs answer to different linguistic types of questions. This research design was motivated by the fact that humans use language models as plug-and-play tools, without any finetuning methods. Future research using these methods might improve vanilla LLM's performance for the QA task. Moreover, LLM-as-a-Judge evaluation could be explored and compared with existing metrics.

Ethical Considerations

The dataset used for this experiments has open license (CC BY-NC 4.0) and can be used for research by the NLP community. The RefoMed-EN dataset contains no personal data or patient data.

Limitations

This work was conducted only on English using only fully open language models, OLMo (Groeneveld et al., 2024) and Pythia (Biderman et al., 2023). Due to computational power limitations, we conducted our analysis on Small Language Models (1b and 7b). Future studies could include more recent models, like OLMo2 (OLMo et al., 2024) and OLMo3 (Olmo et al., 2025), as well as concurrent data tracing tools such as OLMoTrace (Liu et al., 2025). Other open source LLMs give access to their pretraining data, like BLOOM (Workshop et al., 2022) available, but the big size of the dataset makes it difficult to explore.

Moreover, we are aware of this limitation for your semantic similarity evaluation, and we further investigate medical metrics such as MEDCON (Yim et al., 2023), or fact checking metrics and tools like FACTSCORE (Min et al., 2023), FIRE (Xie et al., 2024), or LOKI (Li et al., 2025).

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Ph.D. thesis, Université de Strasbourg.

Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.

Xavier Daull, Patrice Bellot, Emmanuel Bruno, Vincent Martin, and Elisabeth Murisasco. 2023. Complex qa and language models hybrid architectures, survey. *arXiv preprint arXiv:2302.09051*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, and 1 others. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Natalia Grabar and Rémi Cardon. 2018. Clear-simple corpus for medical french. In *ATA*.

Natalia Grabar and Iris Eshkol. 2016. Why do we reformulate? automatic prediction of pragmatic functions. In *HrTAL 2016*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Sanna Iivanainen, Jarkko Lagus, Henri Viertolahti, Lauri Sippola, and Jussi Koivunen. 2024. Investigating large language model (llm) performance using in-context learning (icl) for interpretation of esmo and nccn guidelines for lung cancer.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2401.15269*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Bruce W Lee and Jason Lee. 2023. Lftk: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, and 1 others. 2024a. On the role of long-tail knowledge in retrieval augmented large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–126.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. 2024b. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024c. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, and 1 others. 2025. Olmotrace: Tracing language model outputs back to trillions of training tokens. *arXiv preprint arXiv:2504.07096*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. Medredqa for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Roland Roller, and Sebastian Möller. 2024. Xai for better exploitation of text in medical decision support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513.

Youssra Rebboud, Pasquale Lisena, Lionel Tailhardat, and Raphael Troncy. 2024. Benchmarking llm-based ontology conceptualization: A proposal. In *ISWC 2024, 23rd International Semantic Web Conference*.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Pasi Shailendra, Rudra Chandra Ghosh, Rajdeep Kumar, and Nitin Sharma. 2024. Survey of large language models for answering questions across various fields. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 520–527. IEEE.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Amalia Todirascu, Sebastian Padó, Jennifer Krisch, Max Kisselew, and Ulrich Heid. 2012. French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2024. Fire: Fact-checking with iterative retrieval and verification. *arXiv preprint arXiv:2411.00784*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024b. Knowledgefmath: Knowledge-intensive math reasoning in finance domains. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 12841–12858. Association for Computational Linguistics (ACL).

A RefoMed Linguistic Annotation

The RefoMed (Buhnila et al., 2024) has extensive human annotation validated by human inter-annotator agreement. The annotators were French bilingual speakers of different levels of expertise in linguistics and NLP: senior researchers, PhD and master students specialized. The annotation was conducted on the original version in French and then automatically translated in English with DeepL API. The translation was cleaned and verified, however, some translation mistakes might be scattered in the 6,170 lines dataset.

The annotation comprises a linguistic marking of the lexical relations that exists between the reformulation and the medical term. We show a full example of annotation in Table 9. Different relations were tagged, such as *hyperonymy, hyponymy, meronymy and synonymy*. In the case of pragmatic functions that we explored in this study, the annotators marked the reason to the use of the reformulation in the text: to give a *definition, examples, explanations, denominations* (other name for the same concept), and *paraphrases* of the medical term. RefoMed has annotations of discourse markers that help automatically identify reformulations (all pragmatic types presented above) inside the same sentence, such as "is a", "meaning", "also called", "refers to" (Buhnila, 2023).

B Head and Tail Query Terms from RefoMed-EN

We extract the 50 query terms from head and tail of the Dolma corpus. See Table 6 for 25 terms from expert and general public domain and for each term, the table also includes the number of documents (C_{df}) present in the Dolma corpus. Similarly, we extract 50 terms from tail distribution of Dolma corpus (Table 7 and Table 8).

RefoMed-EN-EX			RefoMed-EN-GP		
Term	DocFreq	TermOcc	Term	DocFreq	TermOcc
life	497066067	13012	professional	161724849	6894
control	230950452	12218	version	159636360	15503
intervention	23396121	5022	surgery	37481694	8558
conditions	136013609	4499	conditions	136013609	4499
function	114612170	14331	skills	126936290	8171
Treatment	113612643	6963	treatment	113612643	6863
Pain	80378923	24840	tool	92230908	13708
patients	78742642	5985	condition	88989985	8446
disease	75724477	9078	scale	88228556	12443
'client'	65025830	9365	long term	83709724	3505
behavior	64089027	6194	Pain	80378923	24840
tests	57368874	21505	disease	75724477	9078
Cancer	50918098	9858	foot	67911272	8771
drugs	44267374	11980	volume	67588213	5281
Cells	43057710	14193	exercise	64859465	7467
lifestyle	42983527	4540	Iron	35692516	11691
MS	38326252	17530	brain	60468547	7769
Surgery	37481694	8558	tests	57368874	21505
Anxiety	35107524	7560	drug	57092723	25334
diseases	34543764	3966	cancer	50918098	9858
IM	33368643	29766	technique	48805868	4938
protein	32887735	17620	procedure	47808948	21328
depression	30980124	9459	symptoms	44391219	5593
infection	27670132	6022	drugs	44267374	11980
facilitate	27181783	2858	MS	38326252	17530

Table 6: Head-50 query terms from RefoMed-EN from the Dolma corpus. We show 25 terms from RefoMed-EN-EX and 25 terms from RefoMed-EN-GP.

C Model Size and Budget

Experiments were done on a P100 NVIDIA GPU, for an individual runtime of $\leq$ 15 hours.

RefoMed-EN-EX		
Term	**DocFreq**	**TermOcc**
Horton's disease and Takayasu's arteritis	1	2
vesico-sphincter dyssynergia (VSD)	1	1
Motor Neurone Disease (MMN)	1	2
strain CF7968	1	1
region of around 100bp	1	1
inappropriate secretion of HAD	1	1
surgical haemorrhagic shock	1	1
psoriatic onychosis	1	3
extracellular proliferating bacteria	1	1
medical problems contributing to insomnia	1	1
pain resulting from a medical cause	1	1
Early administration of an amino acid solution	1	1
recent global weakening	2	2
Crohn's disease and hemorrhagic rectocolitis	2	2
proteins EWI-2 and CD81	2	2
Decompression sickness (DD)	2	6
low risk of bias threshold	2	2
Steinert's myopathy	3	6
cognitive disorders in Parkinson's disease without dementia	2	4
lymphocytic bacterial meningitis	2	3
forms of neurolupus	2	2
family of antithrombotics	2	2
dopamine stock solutions	2	2
CoA HMG reductase inhibitors	2	2
Jet lag and shift work syndromes	2	4

Table 7: Tail-25 query terms from RefoMed-EN-EX based on the Dolma corpus.

RefoMed-EN-GP		
Term	**DocFreq**	**TermOcc**
treatment of ovarian cancer by surgery	1	1
Pseudo-rhizomelic polyarthritis	1	1
Sydenhan's chorea	1	2
Uterine (uterine/endometrial) cancer	1	10
Work-related musculoskeletal disorders of the upper limbs and neck	1	2
reformulation of a food product	1	1
bronchiolitis obliterans syndrome and lymphocytic bronchitis	2	1
ImmunoGuard(R)	2	2
bronchiolitis obliterans syndrome and lymphocytic bronchitis	1	2
prophylaxis with immunoglobulin g	1	1
notion of therapeutic education	1	4
disabled child's education allowance	2	2
Blood Form Count	2	2
rare bullous skin diseases	1	1
psychiatric disorders and social anxieties	1	1
pain resulting from a medical cause	1	1
gushing liquid diarrhea	1	1
disorders of immediate memory	1	1
shiftworker sleep disorder	1	1
disorder of ocular refraction	1	2
Eulenburg's paramyotonia congenita	2	2
notion of therapeutic education	2	4
facio-truncular obesity	2	2
STBBI, STD	2	11
American mucocutaneous and cutaneous leishmaniasis	2	4

Table 8: Tail-25 query terms from RefoMed-EN-GP based on the Dolma corpus.

Context (C)	Myasthenia is an autoimmune disease triggered by autoantibodies that most often target the nicotinic acetylcholine receptor.
Term (T)	Myasthenia
Marker (M)	is a
Indicator (I)	disease
Reformulation (R)	an autoimmune disease triggered by autoantibodies that most often target the nicotinic acetylcholine receptor.
Lexical relation (LR)	hyperonymy
Pragmatic function (PF)	definition
C	Diabetic kidney disease (DKD), also known as diabetic nephropathy (DN) [...]
T	Diabetic kidney disease (DKD)
M	also known as
I	disease
R	diabetic nephropathy (DN)
LR	synonymy
PF	denomination
C	Antibiotics (chloramphenicol, tetracycline and doxycycline) are used in the treatment of this disease.
T	Antibiotics
M	()
I	disease
R	chloramphenicol, tetracycline and doxycycline
LR	hyponymy
PF	exemplification
C	We planned to include all RCTs and quasi-RCTs (in which assignment was not random, but was supposed to be unbiased, e.g., alternate assignment) involving treatment for paraneoplastic neuropathies.
T	quasi-RCTs
M	()
I	for example
R	in which assignment was not random, but was supposed to be unbiased
LR	meronymy
PF	explanation
C	We accepted as comparison arms a placebo control (sham intervention) or without injection, or another active treatment [...]
T	placebo control
M	()
I	intervention
R	sham intervention
LR	synonymy
PF	paraphrase

Table 9: Examples of the RefoMed fine-grained linguistic annotation for each lexical relation and pragmatic function used to determine the question types.

Discharge Instructions are not One Task: Grounding Differences Between Surgical and Non-Surgical Admissions

Mayank Jobanputra[1] **Justin Xu**[2] **Samarth H. Oza**[3] **Hulma Naseer**[1]
Yifan Wang[1] **Blerta Veseli**[1] **Chandralekha Kona**[3] **Haochen Cui**[2]
David W. Eyre[2] **Vera Demberg**[1]

[1]Saarland University [2]University of Oxford [3]Independent Researcher

Abstract

Discharge instructions are patient-facing, safety-critical documents that guide medication use, follow-up care, and recovery after hospitalization. Because they must synthesize information across the clinical record and often include post-discharge guidance not stated verbatim in the EHR, they are a difficult target for clinical text generation. In this work, we study discharge instructions in MIMIC-IV through a grounding-first lens. Using two LLMs, we decompose each discharge instruction into medically relevant statements and verify them against the Electronic Health Record (EHR). We find that discharge instructions for Surgical admissions are much longer, averaging roughly 24–25 statements per admission versus 11–12 in Non-Surgical cases, while supported content remains similar in absolute count. The additional Surgical content is dominated by statements that are not directly stated in the record or require clinically plausible extrapolation. Through this analysis, we advocate for better grounding and completeness evaluations at a fine-grained level, establishing a foundational step toward safer and more reliable discharge-instruction generation.

1 Introduction

Discharge instruction generation has become a central clinical text generation problem because it sits directly at the interface between hospital care and patient self-management. At the center of this progress is the BioNLP 2024 shared task *Discharge Me!*, which established a public benchmark and made discharge documentation a focal task for the community (Xu et al., 2024).

At a high level, discharge instructions should answer three questions for the patient: what happened during this admission, what they should do after leaving the hospital, and what signs should prompt them to seek further medical care. Yet, Discharge instructions can vary a lot depending on the type of admission. Some admissions, especially surgical admissions, contain substantially more content for post-procedural counseling and clinician-authored guidance that may not be directly inferable from the EHR alone. Such cases pose a more challenging grounding problem than shorter, more routine non-surgical admissions. Figure 1 provides an overview of these differences per admission type.

This difference in complexity is easy to overlook in aggregate evaluation. Xu et al. (2024) reports high correctness scores for top-performing systems during clinician review, but it does not provide explicit quantitative detail about the length or procedural complexity of the reviewed discharge instructions. Without that information, it is difficult to know how well current systems handle the most complex and weakly grounded cases. We therefore argue that before asking whether systems can generate discharge instructions well, we should first ask a more foundational question related to grounding and completeness:

> *What content in a discharge instruction can be derived from the EHR?*

In this work, we systematically study the clinician-written discharge instructions themselves to understand which parts of human-written discharge instructions are explicitly grounded in the EHR. Carrying out this study manually is not feasible due to the required number of clinician hours. Hence, we use Large Language Models (LLMs) to analyze the MIMIC-IV dataset, similar to the VeriFact (Chung et al., 2025) framework. For the systematicity, we segment the data into Surgical and Non-Surgical subsets to answer:

1. What are the quantitative differences in the discharge instructions between Surgical and Non-Surgical cases?

2. Which category relies more heavily on ungrounded clinical extrapolation?

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 426–434
July 3-4, 2026 ©2026 Association for Computational Linguistics

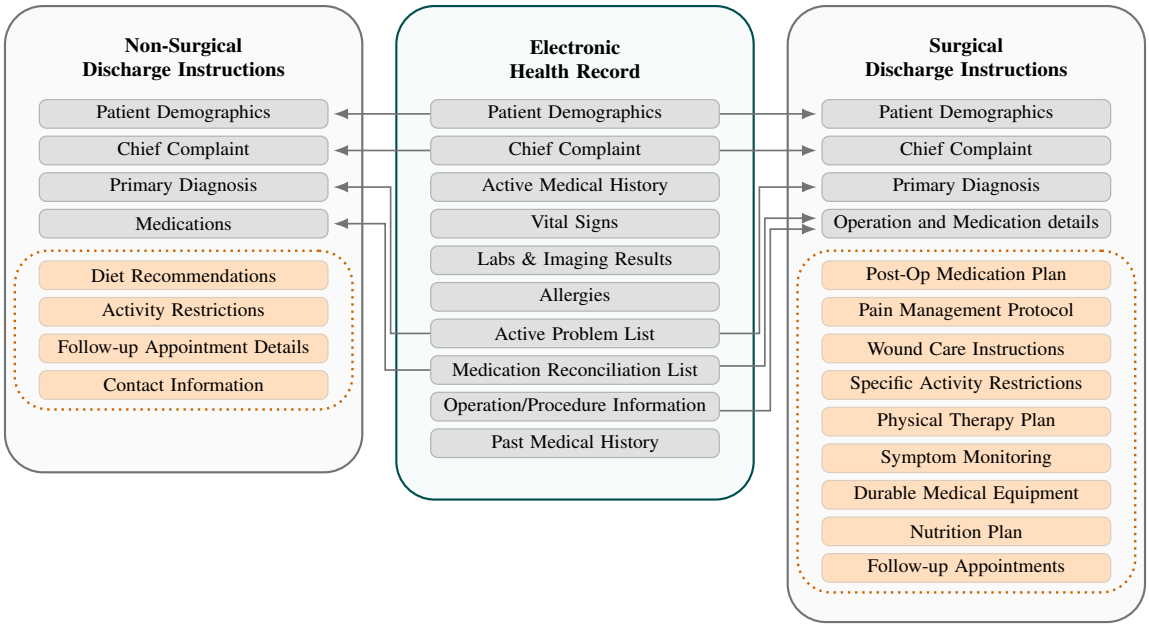


Figure 1: **Surgical and Non-Surgical discharge instructions differ not only in length but in verifiability.** Surgical discharge instructions contain substantially more medically important post-discharge care statements per admission than Non-Surgical instructions. However, the increase is driven primarily by statements that are not directly verifiable from the Electronic Health Records.

Our choice to classify cohorts using the clinical narrative is driven by the structural limitations of structured Electronic Health Record (EHR) fields. While ICD-10 diagnosis codes indicate clinical conditions, they do not reliably confirm whether an operative procedure was performed during a specific hospitalization. Conversely, while procedure codes provide a more direct signal, individual admissions frequently contain multiple overlapping codes. Constructing a rule-based mapping table to resolve these cases would require extensive exception logic and sensitivity analyses to handle procedural ambiguities. Utilizing the text-based clinical narrative allows the pipeline to dynamically determine the primary treatment pathway based on the comprehensive hospital course.

We view this work as a stepping stone toward creating a fine-grained benchmark for evaluating the faithfulness and completeness of generated discharge instructions.

2 Related Work

Early work on discharge-instruction generation used more structured formulations. J Kurisinkel and Chen (2019) cast the task as set-to-ordered-text generation from ICD codes, showing that discharge instructions often require inference rather than direct copying from the input. Re^{3}Writer moves closer to record-grounded generation with a retrieve-reason-refine framework designed to improve faithfulness in patient-instruction generation (Liu et al., 2022). Recent work has focused on benchmarking and improving discharge-section generation at scale. The BioNLP 2024 shared task *Discharge Me!* established a public benchmark for generating *Brief Hospital Course* and *Discharge Instructions*, and representative system papers explored retrieval, section selection, and prompt- or fine-tuning-based methods (Xu et al., 2024; Koontz et al., 2024; Lyu et al., 2024; Damm et al., 2024).

Building upon these evaluations, our work takes a step further by studying the fine-grained clinical differences within the dataset. Specifically, we explore the distinction between Surgical and Non-Surgical admissions to evaluate how grounding complexity varies across different types of patient care.

3 Method

In this section, we provide an overview of each stage of our pipeline for classifying the parts of the discharge instructions.

Dataset. We analyze the *Discharge Me!* Phase II Test split for all our experiments to reduce the risk that the evaluated discharge instructions were seen by the LLMs during instruction tuning or post-training. As the final held-out test split, it provides

a cleaner setting for studying record-grounded verification than training or validation data. This split contains discharge instructions for 10,962 Hospital admissions.

3.1 DI extraction and cleaning

We first transform the discharge notes into a structured sentence-level format using a constrained extraction pipeline. We use GPT-OSS-120B (OpenAI, 2025) to: (i) identify the *Discharge Instructions* section, (ii) copy them verbatim, (iii) remove formatting artifacts and rhetorical questions, (iv) group related clauses into meaningful declarative sentence units, and (v) predict a binary label, SURGICAL or NON-SURGICAL, for the given datapoint. The exact prompts are available in the appendix D.

We observed that GPT-OSS-120B classified most discharge instructions as Non-Surgical. To mitigate this issue, we introduced a secondary deterministic validation step that is applied to the raw clinical narrative. Any admission initially classified as Non-Surgical is re-categorized as Surgical if the discharge instruction documents a major invasive procedure or matches a predefined lexicon of surgical keywords (e.g., *post-op*, *incision*, *sutures*).

3.2 Sentence categorization with LLMs

We then evaluate each discharge-instruction sentence using two LLMs: Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and MediX-R1 (Mullappilly et al., 2026). We choose a Llama-family model because VeriFact uses Llama 3.1 70B for EHR-grounded verification, and we include MediX-R1 as a recent medically specialized LLM with strong reported benchmark performance. Each LLM decomposes a larger declarative sentence unit into smaller medically important statements, retrieves supporting or refuting evidence from the diagnoses and hospital course, and assigns a verification label. The exact prompts are available in the appendix D. We use three main verification labels: *Supported*, *Not Supported*, and *Not Addressed*.

Supported statements are explicitly backed by the clinical context. *Not Supported* statements are addressed by the record but only partially supported, lack sufficient detail, or are inconsistent with it. *Not Addressed* statements are not directly verifiable from the available evidence. We also track two finer subcategories. Within *Not Supported*, we separate *Contradiction* from statements that lack sufficient detail. Within *Not Addressed*, we separate *Clinically Plausible Extrapolation* from statements that are completely not addressed by the record. Since, we are evaluating clinician written summaries, we expect most of the statements to be either *Supported* or *Clinically Plausible Extrapolation*. Other cases may reveal the limitations of LLMs or genuine errors by clinicians while writing Discharge Instructions.

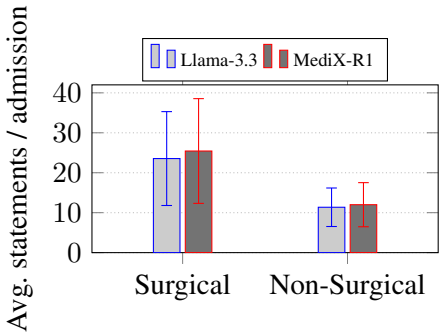


Figure 2: Average number of medically important statements per admission, with standard deviation as error bars. Surgical discharge instructions are approximately twice as long as Non-Surgical ones for both LLMs.

Code. Our code is available on GitHub[1].

4 Results

We start by analyzing the lengths of discharge instructions. We find that the discharge instructions for Surgical admissions are much longer than Non-Surgical admissions. Figure 2 shows that this difference is large and consistent across both LLMs. This suggests that the grounding problem is not uniform across the dataset. Surgical discharge instructions place a substantially larger informational burden on any generation or verification system.

Figure 3 shows that the additional content in Surgical discharge instructions is concentrated in statements that are not directly verifiable from the available EHR evidence (see Table 4 in the Appendix for exact details). For both LLMs, the number of *Supported* statements is similar in Surgical and Non-Surgical admissions, but Surgical admissions contain many more statements in the broad *Not Addressed* category. This means that Surgical discharge instructions are longer not because they contain more directly supported information, but because they include more content that is missing from the diagnoses and Brief Hospital Course or requires inference. Surprisingly, within the *Not*

[1] https://github.com/mayankjobanputra/dischage_differences

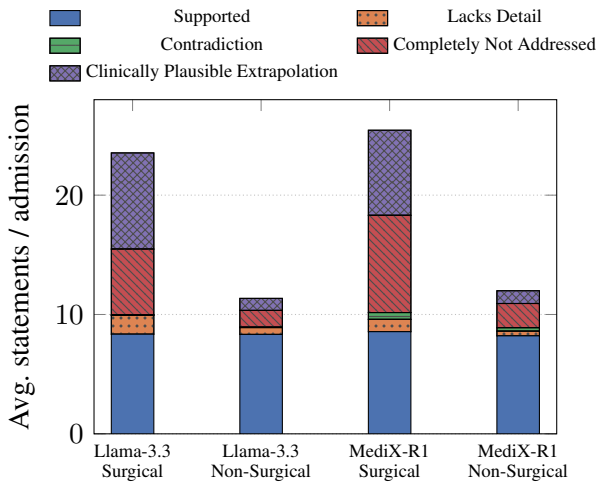


Figure 3: Fine-grained classification results for Surgical and Non-Surgical hospital admissions.

Category	Surgical	Non-Surgical
Broad labels (%)	79.79	85.16
Broad-label κ	0.6366	0.6645
Fine labels (%)	66.02	81.02
Fine-label κ	0.4845	0.5955

Table 1: Agreement stats for Llama-3.3 and MediX-R1. Broad labels refer to *Supported*, *Not Supported*, and *Not Addressed*; fine labels use the full five-label scheme.

Addressed category, the increase in Surgical admissions is distributed across both *Completely Not Addressed* statements and *Clinically Plausible Extrapolation*. We discuss our findings from the manual error analysis on this in Section 5.

Next, we measure the agreement between the two LLMs. Overall, we observe over 82% agreement (κ = 0.66) between Llama-3.3 and MediX-R1 at the broad categories level, but at the fine-grained level, the agreement drops to 72.41% (κ = 0.53). We provide a more detailed analysis in Table 1.

Based on the results from Table 1, we can also see that the agreement is lower in the Surgical admissions than in the Non-Surgical admissions at both the broad level and the fine level. While the agreement levels are encouraging, the overall results need to be verified by clinicians before they can be taken as evidence that the LLM labels are reliable on their own and also to decide which LLM performs better on this classification task.

5 Human Evaluation

To test whether the LLM labels are clinically meaningful, we conducted a targeted human study on 24 admissions comprising 544 medically important statements. The study was stratified across the three subsets introduced above: *Cross-LLM Consensus*, *Cross-LLM Disagreement*, and *Clinically Plausible*. One clinician annotated all 24 admissions. A second clinician independently annotated a subset of 6 admissions, yielding 146 aligned statements for inter-rater agreement analysis.

Comparison	Agreement (%)	κ
Human–human (broad)	96.58	0.8874
Human–human (fine)	91.78	0.7785
Human–Llama-3.3 (broad)	81.2	0.4819
Human–Llama-3.3 (fine)	45.8	0.2333
Human–MediX-R1 (broad)	71.1	0.3591
Human–MediX-R1 (fine)	45.6	0.1918

Table 2: Human-evaluation results. Human–human agreement is computed on the double-annotated subset of 146 aligned statements. Human–Llama-3.3 and Human–MediX-R1 agreements are computed on the 544 aligned statements.

Table 2 shows high human–human agreement at both the broad level ($\kappa = 0.8874$) and the fine level ($\kappa = 0.7785$), indicating that the labeling scheme can be applied consistently by clinicians.

At the broad level, Llama-3.3 agrees with the human annotations more than MediX-R1 (81.2% vs. 71.1%; $\kappa = 0.4819$ vs. 0.3591). At the fine level, however, both LLMs perform similarly poorly (45.8% vs. 45.6%), with low κ values in both cases. This shows that neither LLM is reliable enough to replace clinician judgment, especially for closely related categories such as *Not Addressed* and *Clinically Plausible Extrapolation*.

Error analysis The most frequent error is confusion between *Not Addressed* and *Clinically Plausible Extrapolation*. Routine post-operative advice, such as lifting restrictions, driving restrictions while taking pain medication, light exercise, and incision care, is often labeled *Not Addressed* by the LLMs even when the clinician labels it *Clinically Plausible Extrapolation*.

A second recurring error is over-attribution of support to discharge boilerplate. MediX-R1 more often labels service-level instructions as *Supported*, while the clinician more often labels them *Clinically Plausible Extrapolation* or *Not Addressed*. Examples include refill instructions, brace use, diet advice, and resumption of home medications.

MediX-R1 appears relatively stronger on procedure-specific *Clinically Plausible Extrapolation*, while Llama-3.3 is stronger on explicitly

Supported statements and achieves higher overall agreement with human labels. We provide example statements along with human and LLM labels in Appendix C.

6 Conclusion

In this work, we show that Surgical and Non-Surgical admissions differ substantially in both length and verifiability: Surgical discharge instructions contain many more medically important statements, and the increase is driven mainly by content that is not directly supported by the EHR text. This suggests that future work on discharge-instruction generation should report results separately for Surgical and Non-Surgical cohorts and move toward more fine-grained evaluation of grounding. Our human study shows that LLMs can help surface likely ungrounded content and identify difficult regions of the dataset, but they are not reliable enough for fine-grained labeling without clinician validation. Based on our findings, we suggest that reliable evaluation of generated discharge instructions likely requires resources beyond the EHR alone. A potential future direction can be an agentic setup that can use additional clinical tools, protocols, and knowledge sources in a Biomni-like framework (Huang et al., 2025) adapted to discharge-specific needs.

Limitations

Our study has some limitations. We use only open-weight LLMs in our experiments and do not compare against strong closed-source models, which may differ in both decomposition quality and verification behavior. We also report results from a single run per model, so we do not measure variability across random seeds or decoding configurations. Our analysis is based on a single prompt configuration per stage and does not explore a broader prompt sensitivity study. Different prompt wording, output constraints, or evidence-formatting choices could affect both the extracted statements and the assigned verification labels. More generally, we evaluate only two LLMs, so the observed agreement patterns may not fully characterize the behavior of other model families. We also do not provide any analysis on how exactly this affects recent LLMs's capability of generating factual and clinically complete discharge instructions.

Ethics Statement

This work studies de-identified clinical text and proposes auditing methods for safety-critical patient-facing generation. We emphasize that unsupported or weakly supported discharge instructions can create direct patient harm, and that automated systems should not be deployed without clinician oversight, calibration, and site-specific validation. We also do not use any closed-source LLMs such as GPT-5, Gemini to avoid any kind of sensitive data leakage. All clinicians participating in the study are board-certified physicians and are co-authors of this work.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and insightful comments. MJ and VD acknowledge the funding by Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science. YW and BV were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914. JX gratefully acknowledges joint support from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). HC is supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) Grant Number 2024-I2M-2-001-1.

References

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1):AIdbp2500418.

Hendrik Damm, Tabea Margareta Grace Pakull, Bahadır Eryılmaz, Helmut Becker, Ahmad Idrissi-Yaghir, Henning Schäfer, Sergej Schultenkämper, and Christoph M. Friedrich. 2024. WisPerMed at "discharge me!": Advancing text generation in healthcare with large language models, dynamic expert selection, and priming techniques on MIMIC-IV. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 105–121, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Litton J Kurisinkel and Nancy Chen. 2019. Set to ordered text: Generating discharge instructions from medical billing codes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6165–6175, Hong Kong, China. Association for Computational Linguistics.

Jordan C. Koontz, Maite Oronoz, and Alicia Pérez. 2024. Ixa-med at discharge me! retrieval-assisted generation for streamlining discharge documentation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 658–663, Bangkok, Thailand. Association for Computational Linguistics.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David A. Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Advances in Neural Information Processing Systems*, volume 35.

Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. UF-HOBI at "discharge me!": A hybrid solution for discharge summary generation through prompt-based tuning of GatorTronGPT models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 685–695, Bangkok, Thailand. Association for Computational Linguistics.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. 2026. Medix-r1: Open ended medical reinforcement learning. *Preprint*, arXiv:2602.23363.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset statistics

The summary of corpus and label is provided in Table 3.

B Detailed Distribution Statistics

Table 4 provides the numerical distribution of the broad and fine-grained verification categories for both Llama-3.3 and MediX-R1.

C Qualitative Error Analysis

The examples in Table 5 summarizes the main qualitative error patterns observed in the human evaluation. These errors are grouped by the subset from which they were sampled.

D Prompt Templates

D.1 DI extraction and cleaning: system prompt

```
You are an expert clinical data extraction
↪  assistant. Your primary directive is to
↪  analyze complex hospital admission
↪  records and extract specific,
↪  structured information into a strictly
↪  formatted JSON object.
```

D.2 DI extraction and cleaning: task prompt

```
**TASKS:**

**Task-1: Verbatim Copy of Discharge
↪  Instructions**
- Locate the "Discharge Instructions"
↪  section within the clinical note.
- Extract and copy the entire text of this
↪  section verbatim, exactly as it appears
↪  in the text.

**Task-2: Split and Group into Meaningful
↪  Sentences**
- Take the verbatim discharge instructions
↪  and split them into declarative
↪  sentences.
- Remove all rhetorical or conversational
↪  questions (e.g., "Questions?").
- Remove all splitting markers or visual
↪  separators like "========" or "****".
- **IMPORTANT:** You must group sentences
↪  that talk about the exact same topic or
↪  refer to each other into a single
↪  string item.
  - *Example of Grouping:* If the text says:
  ↪  "We have also prescribed a nebulizer
  ↪  machine that you can use at home. We
  ↪  only recommend using this if you are
  ↪  too short of breath to use your
  ↪  albuterol inhaler effectively."
    *Do NOT split this. Output it as a
    ↪  single array item:* ["We have also
    ↪  prescribed a nebulizer machine that
    ↪  you can use at home. We only
    ↪  recommend using this if you are too
    ↪  short of breath to use your
    ↪  albuterol inhaler effectively."]

**Task-3: Identify the Treatment Service
↪  Type**
```

	Llama-3.3	MediX-R1
Corpus statistics		
Admissions with outputs from both models	10,958	
Total sentences	74,576	77,281
Sentences / admission (avg.)	6.81	7.05
Sentences / admission (max.)	53	49
Total Medical statements	159,280	168,649
Medical statements / admission (avg.)	14.54	15.39
Medical statements / admission (max.)	102	115
Verification label counts		
Supported	94,669	93,503
Not Supported	8,571	5,793
Lacks Detail	8,060	1,936
Contradiction	511	3,857
Not Addressed	55,529	65,496
Completely Not Addressed	26,144	38,627
Clinically Plausible Extrapolation	29,385	26,869

Table 3: Summary of corpus and label statistics.

	Surgical Admissions		**Non-Surgical Admissions**	
Category	**Llama-3.3**	**MediX-R1**	**Llama-3.3**	**MediX-R1**
Broad Labels				
Supported	8.35 ± 5.58	8.56 ± 6.08	8.33 ± 3.96	8.21 ± 4.06
Not Supported	1.64 ± 2.32	1.59 ± 2.13	0.61 ± 0.93	0.69 ± 1.02
Not Addressed	13.55 ± 9.61	15.29 ± 10.43	2.41 ± 2.27	3.10 ± 2.89
Fine-Grained Subcategories				
Lacks Detail	1.57 ± 2.28	1.03 ± 1.70	0.57 ± 0.88	0.39 ± 0.78
Contradiction	0.07 ± 0.33	0.57 ± 1.24	0.04 ± 0.25	0.29 ± 0.61
Completely Not Addressed	5.49 ± 7.00	8.15 ± 8.16	1.40 ± 1.57	2.03 ± 1.77
Clinically Plausible Extrapolation	8.06 ± 8.60	7.13 ± 8.11	1.01 ± 1.74	1.07 ± 2.16

Table 4: Mean and standard deviation (Mean $\pm$ SD) of medically important statements per admission assigned by Llama-3.3 and MediX-R1 across all evaluation categories.

```
- Analyze the hospital course and diagnoses.
- Choose `SURGICAL` if the patient underwent
↪  a major therapeutic surgical procedure
↪  in an operating room during this stay
↪  and the "Service" is "SURGERY".
- Choose `NON-SURGICAL` if the patient was
↪  managed medically.
*(Note: Try your best to classify this
↪  based on the clinical narrative. The
↪  system will double-check your work).*

**Task-4: Extract Major Surgical
↪  Procedures**
- If the service is `SURGERY`, list the
↪  names of the "Major Surgical or
↪  Invasive Procedure" performed.
- If no major surgical procedure is found or
↪  the service is `MEDICINE`, return
↪  ["NA"].

**Task-5: Anonymization**
- Assign a fake name to the patient
↪  `Noname`.
- Ensure any remaining anonymization
↪  markers (like `___` or `[**...**]`) are
↪  replaced with generic placeholder
↪  `[masked info]`.

**REQUIRED JSON OUTPUT FORMAT:**
{
  "patient_name": "<Output from Task 5>",
  "verbatim_discharge_instructions":
  ↪  "<Output from Task 1>",
  "discharge_instructions_sentences":
  ↪  ["<Output from Task 2>"],
  "clinical_pathway": "<Output from Task
  ↪  3>",
  "procedures": ["<Output from Task 4>"]
}

**PATIENT DATA:**
Subject ID: {{ subject_id }}
Hospital Admission ID: {{ hadm_id }}

--- DIAGNOSES ---
{{ diagnoses_text }}

--- RAW CLINICAL NOTE ---
{{ raw_note_text }}
```

Subset	Claim	Human	Llama	Medix
Cross-LLM Consensus	"Do not drive until you have stopped taking pain medicine and feel you could respond in an emergency"	CPE	NA	NSup.
Cross-LLM Consensus	"Don't lift more than [masked info] lbs for 4 weeks"	CPE	NA	NA
Cross-LLM Consensus	"You may start some light exercise when you feel comfortable"	CPE	NA	NA
Cross-LLM Consensus	"You will need to stay out of bathtubs or swimming pools for a time while your incision is healing"	CPE	NA	NA
Cross-LLM Consensus	"You may feel weak or washed out for a couple of weeks"	CPE	NA	NA
Cross-LLM Disagreement	"You may climb stairs"	CPE	NA	Sup.
Cross-LLM Disagreement	"Eat a normal healthy diet."	CPE	NSup.	Sup.
Cross-LLM Disagreement	"You do not need a brace."	CPE	NA	Sup.
Cross-LLM Disagreement	"You should resume taking your normal home medications."	CPE	NSup.	Sup.
Cross-LLM Disagreement	"Please allow 72 hours for refill of narcotic prescriptions, so please plan ahead."	CPE	NA	Sup.
Cross-LLM Disagreement	"We are not allowed to call in narcotic prescriptions to the pharmacy."	NA	NA	Sup.
Clinically Plausible	"You may have a sore throat because of a tube that was in your throat during surgery"	CPE	NA	CPE
Clinically Plausible	"You may shower at this time but keep your incision dry."	CPE	NA	CPE
Clinically Plausible	"It is best to keep your incision open to air but it is ok to cover it when outside."	CPE	NA	CPE
Clinically Plausible	"Avoid heavy lifting, running, climbing, or other strenuous exercise until your follow-up appointment."	CPE	NA	CPE
Clinically Plausible	"You may take leisurely walks and slowly increase your activity at your own pace once you are symptom free at rest."	CPE	NSup.	CPE
Clinically Plausible	"No contact sports until cleared by your neurosurgeon."	CPE	NA	CPE

Table 5: Representative qualitative disagreement cases from the human evaluation. CPE = *Clinically Plausible Extrapolation*, NA = *Not Addressed*, NSup. = *Not Supported*, Sup. = *Supported*.

D.3 Sentence categorization: system prompt

```
You are an expert attending physician and clinical auditor. Your task is to verify the factual accuracy of hospital discharge instructions against the patient's ground-truth clinical record.
```

D.4 Sentence categorization: evaluation prompt

```
**TASKS:**

**Task-1: Atomic Extraction**
- For each sentence in the "TARGET SENTENCES TO EVALUATE" list, break it down into independent, verifiable atomic claims.
- Example: "Take 5mg of Lisinopril daily and follow up in 2 weeks." -> Claim 1: "Take 5mg of Lisinopril daily.", Claim 2: "Follow up in 2 weeks."

**Task-2: Evidence Retrieval and Reasoning**
- Search the provided CLINICAL CONTEXT for evidence that supports or refutes each atomic claim.
- Extract the direct quotes from the context (`evidence_quoted`). Write "None" if the claim is not addressed.
- Explain your clinical logic step-by-step (`thought_process`) *before* assigning a verdict.
- **CRITICAL JSON RULE:** Do NOT use double quotes (") inside your `thought_process`, `atomic_claim`, or `evidence_quoted` fields. If you must quote something from the text, use single quotes ('). Unescaped double quotes will corrupt the JSON object.

**Task-3: Verdict Classification**
- Classify each atomic claim into exactly one of the following categories (`verdict`):
  - `Supported`: The claim is explicitly backed by the context.
  - `Clinically Plausible Extrapolation`: The claim isn't explicitly written, but is standard, safe medical practice based on the diagnoses (e.g., advising a heart failure patient to limit salt).
  - `Contradiction`: The claim directly conflicts with the context (e.g., wrong medication, wrong dosage, wrong follow-up timeline).
  - `Not Addressed`: The claim cannot be verified or safely extrapolated from the context.
  - `Not Supported`: If the claim is addressed and not fully supported by the reference context. This includes cases where the text is only partially supported by the reference context.
```

```
**REQUIRED JSON OUTPUT FORMAT:**
{
  "evaluations": [
    {
      "original_sentence": "<The full target
      ↪  sentence being evaluated>",
      "atomic_evaluations": [
        {
          "atomic_claim": "<The extracted
          ↪  atomic claim>",
          "thought_process": "<Your
          ↪  step-by-step reasoning>",
          "evidence_quoted": "<Direct quote
          ↪  from context or 'None'>",
         "verdict": "<Supported | Clinically
          ↪  Plausible Extrapolation |
          ↪  Contradiction | Not Addressed
          ↪  | Not Supported>"
        }
      ]
    }
  ]
}

**CLINICAL CONTEXT:**
--- Diagnoses ---
{{ diagnoses_text }}

--- Hospital Course ---
{{ hospital_course_text }}

**TARGET SENTENCES TO EVALUATE:**
{{ sentences_json_list }}

Return the result strictly as a structured
↪  JSON object matching the requested
↪  schema.
```

PrionNER: A Named Entity Recognition Dataset for Prion Disease Biomedical Literature

An Dao[1,5*] **Nhan Ly**[2*] **Thao Tran**[2*] **Yuji Matsumoto**[3,4] **Akiko Aizawa**[5]
[1]The University of Tokyo, Tokyo, Japan
[2]Medical Doctor, Independent Researcher
[3] Center for Language AI Research, Tohoku University, Sendai, Japan
[4] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[5] National Institute of Informatics, Tokyo, Japan
dtan@g.ecc.u-tokyo.ac.jp, trinhanly1996@gmail.com, thaotran1490@gmail.com,
yuji.matsumoto.a4@tohoku.ac.jp, aizawa@nii.ac.jp

Abstract

Prion diseases are rare, rapidly progressive, and fatal neurodegenerative disorders that remain difficult to diagnose, particularly in their early stages because of nonspecific clinical presentations. However, to our knowledge, there is no publicly available prion-disease-focused dataset designed to capture a broad range of clinically relevant entities from the biomedical literature. We introduce **PrionNER**, a manually annotated named entity recognition dataset for prion disease clinical information in PubMed abstracts. The current release comprises 317 abstracts, 2,943 sentences, and 6,955 text-bound entity annotations spanning 15 coarse-grained and 31 fine-grained clinically oriented entity types covering diseases, symptoms, diagnostics, findings, anatomy, treatments, and temporal and statistical evidence. Inter-annotator agreement reaches 81.78 exact-match F1, indicating strong annotation consistency. We benchmark supervised BERT baselines, W2NER, and zero-shot extractors on PrionNER. W2NER is the strongest supervised model, and Gemma-4-31B is the strongest zero-shot model, but the benchmark remains challenging, especially for structurally complex mentions and fine-grained clinically adjacent label distinctions. PrionNER provides a clinically grounded benchmark for prion-disease information extraction and supports research on rare-disease biomedical NLP under low-resource, fine-grained, and non-flat extraction conditions. The dataset, annotation guidelines, and evaluation scripts are available at https://github.com/daotuanan/PrionNER/.

1 Introduction

Prion diseases are rapidly progressive, fatal neurodegenerative disorders caused by the misfolding and accumulation of pathological prion protein (Prusiner, 1998). They remain untreatable while carrying a severe risk of iatrogenic transmission (Centers for Disease Control and Prevention, 2026). Early diagnosis is difficult because these diseases are rare and therefore unfamiliar to many general physicians, while their initial symptoms are often vague and overlap with other psychiatric and neurological conditions (Geschwind, 2015). By the time the disease reaches a clearly recognizable stage, it is often already terminal, limiting opportunities for intervention and increasing the risk of inadvertent spread during routine medical procedures (Vallabh et al., 2020). Consequently, recent research efforts have emphasized earlier diagnosis both to reduce transmission risk and to identify patient cohorts for future therapeutic trials (Inada Shimamura and Satoh, 2025). To achieve this, consolidating scattered clinical knowledge into a unified, foundational dataset is essential.

Named entity recognition (NER) has substantially advanced clinical data extraction by structuring key concepts from biomedical text, including diseases, symptoms, diagnostic tests, and biomarkers (Lee et al., 2020; Gu et al., 2021). This progress is supported by widely used biomedical and clinical corpora such as NCBI Disease (Dogan et al., 2014), BC5CDR (Li et al., 2016), CRAFT (Bada et al., 2012), MedMentions (Mohan and Li, 2019), i2b2/VA (Uzuner et al., 2011), and ShARe/CLEF (Pradhan et al., 2013). However, most existing resources target broad biomedical domains, disease mentions, or general classifications rather than the clinically rich diagnostic evidence needed for rare-disease recognition in biomedical literature. Rare-disease-oriented datasets also focus more narrowly on disease, symptom, sign, or phenotype extraction than on the broader diagnostic schema needed for prion disease (Martínez-deMiguel et al., 2022; Groza et al., 2015; Shyr et al.,

*These authors contributed equally to this work.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 435–463
July 3-4, 2026 ©2026 Association for Computational Linguistics

2024). As a result, there is still no publicly available prion-focused NER benchmark that captures the heterogeneous evidence clinicians rely on when distinguishing prion disease and its subtypes from related conditions. To address this gap, we introduce **PrionNER**, a manually annotated dataset for named entity recognition in prion disease clinical narratives derived from PubMed abstracts.

PrionNER contains 317 abstracts with 15 coarse-grained and 31 fine-grained entity types, and its pre-adjudication double-annotated test split reaches 81.78 entity-level exact agreement F1. We evaluate PrionNER using both supervised biomedical encoders and zero-shot extraction models. Among supervised models, W2NER is strongest in both coarse-grained and fine-grained `flat-ner`, reaching 81.86 F1 and 80.46 F1, respectively. Gemma-4-31B is the strongest zero-shot model in `flat-ner`, reaching 71.41 coarse-grained F1 and 68.41 fine-grained F1. However, `non-flat-ner` remains difficult for all models, with the best F1 scores of W2NER (supervised) reaching 13.48 in coarse evaluation and 13.70 in fine evaluation. These results show that PrionNER is learnable but still challenging: performance consistently drops from coarse-grained to fine-grained prediction, and the remaining difficulty reflects a long-tailed label distribution, clinically adjacent type distinctions, and nested or discontinuous mentions.

This work provides a foundation for prion-disease information extraction and clinically oriented biomedical NLP. More broadly, PrionNER illustrates a general challenge in rare-disease biomedical NLP: clinically useful extraction often depends on modeling heterogeneous evidence types, fine-grained diagnostic distinctions, and non-flat mention structures. In this sense, the dataset is relevant beyond prion disease itself and can serve as a compact case study for building clinically grounded resources in other specialized biomedical subdomains.

2 Related Work

In the biomedical literature, JNLPBA (Collier et al., 2004), NCBI Disease Corpus (Dogan et al., 2014), BC5CDR (Li et al., 2016), CRAFT (Bada et al., 2012), and MedMentions (Mohan and Li, 2019) provide benchmark resources for entity recognition, normalization, and concept annotation in MEDLINE or PubMed texts. In the clinical domain, i2b2/VA (Uzuner et al., 2011), ShARe/CLEF (Pradhan et al., 2013), and MedDec (Elgaar et al., 2024) provide corpora from patient reports, clinical notes, and discharge summaries that focus on clinically meaningful entities or decision spans such as problems, tests, treatments, disorder mentions, and medical decisions.

Resources closest to our setting are mainly rare-disease or phenotype-oriented. The HPO corpora (Groza et al., 2015) target phenotype concept recognition and normalization, while RareDis (Martínez-deMiguel et al., 2022) extends this line with annotations for diseases, rare diseases, symptoms, signs, and anaphoric mentions. However, based on our preliminary inspection of the released RareDis annotations, prion-disease coverage remains limited, and these datasets do not provide the broader clinically oriented schema needed for prion-disease information extraction. More broadly, prior resources reflect different tradeoffs in domain breadth and label granularity: corpora such as NCBI Disease and BC5CDR emphasize wider biomedical coverage with a small number of target categories (Dogan et al., 2014; Li et al., 2016), whereas our goal is richer evidence coverage within a single specialized domain.

For prion disease specifically, PDDB (Gehlenborg et al., 2009) is a structured transcriptomic resource rather than a text annotation benchmark. To our knowledge, no publicly available dataset currently targets prion-disease named entity recognition in the biomedical literature with a broad clinically oriented annotation schema. PrionNER fills this gap by providing a manually annotated dataset for prion disease clinical narratives in PubMed abstracts. We next describe how the corpus was collected, filtered, and annotated.

3 Dataset Construction

3.1 Data Sources

We constructed the corpus from PubMed abstracts retrieved with a keyword-based Boolean query over the title and abstract fields; the full query is provided in Appendix Section A.1. The query combined general and subtype-specific prion disease terms, including *Prion Diseases*, *Creutzfeldt-Jakob Disease*, *CJD*, *sporadic CJD*, *familial/genetic CJD*, *variant CJD*, *iatrogenic CJD*, *Kuru*, *Gerstmann-Straussler-Scheinker*, and *Fatal Familial Insomnia* (*FFI*), with clinically oriented terms such as *diagnosis*, *clinical*, *symptoms*, *case*, *progression*, and *treatment*. To bias retrieval toward human clinical

Dataset	Domain	Task	Text Source	Entity Coverage	Size
JNLPBA (Collier et al., 2004)	Biomedical	NER	MEDLINE abstracts	Protein, DNA, RNA, cell line, cell type	2,404 abstracts
I2B2/VA 2010 (Uzuner et al., 2011)	Clinical	NER + RE	Patient reports	Problems, tests, treatments, assertions	871 reports
CRAFT (Bada et al., 2012)	Biomedical	Concept Ann.	Full-text articles	Ontology-linked biomedical concepts	97 articles
SHARE/CLEF (Pradhan et al., 2013)	Clinical	NER + Norm.	Clinical reports	Disorder mentions	300 reports
NCBI DISEASE (Dogan et al., 2014)	Biomedical	NER + Norm.	PubMed abstracts	Disease mentions linked to MeSH/OMIM	793 abstracts
HPO CORPORA (Groza et al., 2015)	Rare disease	NER + Norm.	PubMed abstracts	Phenotype/HPO concept mentions	228 abstracts
BC5CDR (Li et al., 2016)	Biomedical	NER + RE	PubMed abstracts	Chemicals, diseases, chemical-disease links	1,500 abstracts
MEDMENTIONS (Mohan and Li, 2019)	Biomedical	Concept Ann.	PubMed abstracts	UMLS-linked biomedical concepts	4,392 abstracts
RAREDIS (Martínez-deMiguel et al., 2022)	Rare disease	NER + RE	Rare-disease texts	Disease, rare disease, symptom, sign, anaphor	1,041 texts
PrionNER (ours)	Prion disease	NER	PubMed abstracts	15 coarse / 31 fine clinically relevant types	317 abstracts

Table 1: Comparison of PrionNER with representative prior datasets and related resources.

narratives, we excluded terms commonly associated with animal or basic-science studies, including *mice*, *mouse*, *rat*, *animal*, *cell*, *protein*, and *in vitro*. The PubMed query returned 3,414 abstracts, and 3,138 remained after basic preprocessing, including removal of records with empty abstracts.

A pilot manual screening of approximately 500 abstracts by two annotators showed that the retrieved set still contained many off-target papers unrelated to clinical prion disease narratives. Common exclusion cases included basic science, animal or other non-human research, non-clinical analyses such as economic or purely epidemiological studies, and papers in which prion disease was not the main focus. Based on this pilot review, we defined an operational criterion for *related* abstracts and used GPT-5.4 (OpenAI, 2026), prompted as described in Appendix Section A, to screen the 3,138 preprocessed abstracts and reduce manual review workload. Under this criterion, abstracts were retained if they primarily concerned human prion disease in a clinical context, including diagnosis, symptoms, disease progression, or treatment. Across the full 3,138-abstract pool, GPT-5.4 predicted 1,304 abstracts as related and 1,834 as unrelated. We then manually reviewed abstracts from the screened pool together with the pilot-screened set to confirm relevance and remove duplicates. In total, we manually reviewed 1,383 abstracts, including 868 abstracts rated as related by GPT-5.4. Within this set, 772 abstracts were labeled related and 611 were labeled not related by human review. The GPT-5.4 model achieved 90.60 accuracy and 97.80 recall for the related class; it missed only 17 relevant abstracts but incorrectly marked 113 irrelevant abstracts as related, indicating a high-recall but over-inclusive screening strategy. This bias was appropriate for corpus construction because missing clinically relevant rare-disease abstracts would have been more costly than forwarding some extra candidates to manual review. Although we initially intended to annotate all 772 eligible abstracts, practical time constraints limited the current release to 317 annotated abstracts. We therefore treat the present release as a high-quality first benchmark for this specialized domain rather than an exhaustive survey of all eligible prion-disease abstracts. Appendix Section A also provides the relevance-filter audit.

3.2 Entity Schema

We designed the schema to capture clinically meaningful evidence needed for prion-disease diagnostic reasoning and future knowledge graph construction. It was developed through pilot annotation and iteratively revised after joint review of 20 shared annotated abstracts by two medical doctors with experience in neurological diseases together with one coordinator, so that the labels remained clinically meaningful while also remaining coherent from an NLP annotation perspective. Because the dataset is intended to support future clinical decision-support work for prion disease diagnosis, the schema reflects the evidence integration a clinician uses when evaluating a suspected case. Compared with earlier biomedical and rare-disease corpora, which often focus on narrower mention types such as diseases, chemicals, phenotypes, symptoms, signs, or anaphora, our schema is broader because prion diagnosis depends on combining multiple kinds of

diagnostic evidence (Dogan et al., 2014; Li et al., 2016; Groza et al., 2015; Martínez-deMiguel et al., 2022). To reflect this workflow, we organize the schema into three groups: *Case Input*, *Case Diagnosis*, and *Clinical Course and Context*. *Case Input* captures the information available at a patient's presentation, including `Age`, `Symptom`, `Test_name`, `Sequences`, `Anatomic_location`, and `Findings`. *Case Diagnosis* captures the interpretation of those inputs, including disease names, subtype labels (`Generic_Prion`, `Sporadic_Prion`, `Familial_Prion`, `Acquired_Prion`), and alternative conditions considered during evaluation (`Differential_Diagnosis`). *Clinical Course and Context* captures clinically relevant supporting information, including `Treatment`, `Complication`, `Time`, and `Stats`. This broad-entity, focused-domain design differs from prior related datasets such as RareDis, which do not combine the same breadth of clinically oriented entities within a single prion-focused annotation schema (Martínez-deMiguel et al., 2022). We include these categories selectively because of their direct clinical relevance and their potential to support diagnosis in realistic settings. The schema comprises **15** coarse-grained types and **31** fine-grained entity types, and annotation was performed at the mention level using minimal span selection. Table 2 provides a short schema summary for the main text, and Appendix Table 8 provides full definitions with representative examples.

3.3 Annotation Workflow and Quality Control

Pilot annotation. Annotation began with a pilot phase of 20 PubMed abstracts. During this stage, two annotators independently labeled entity mentions in order to identify ambiguities in span boundaries, type assignment, and difficult clinical expressions. Disagreements were reviewed jointly, and these pilot annotations were used to establish the initial annotation guidelines before large-scale annotation. To preserve anonymity, we refer to the two annotators as *Annotator 1* and *Annotator 2*.

Training annotation and guideline refinement. Using the pilot guidelines, the remaining training abstracts were annotated individually, with 151 abstracts annotated by *Annotator 1* and 96 abstracts annotated by *Annotator 2*. During this stage, the guidelines were iteratively refined to address recurring ambiguities and improve consistency across the corpus.

Double annotation, agreement measurement, and final test-set selection. After the training-stage guideline refinement, an additional 70 abstracts were independently annotated by both annotators. Inter-annotator agreement was computed on these independent annotations before any disagreement resolution. Only after agreement measurement did the annotators jointly adjudicate disagreements in this 70-abstract subset to produce a single finalized annotation layer containing 787 sentences and 1,806 text-bound entity annotations. We designated this double-annotated and adjudicated subset as the test split because it is the highest-confidence evaluation subset in the corpus: both annotators labeled these abstracts independently, agreement was measured on the pre-adjudication annotations, and disagreements were then resolved to create the final released test annotations. The clarified adjudication decisions were then propagated back to the training annotations to ensure consistency with the final guideline version, and the final dataset was validated for annotation integrity, including span correctness, label consistency, and the absence of duplicate or conflicting entity annotations.

4 Dataset Statistics

4.1 Corpus Overview

Overall, PrionNER contains 317 abstracts, 2,943 sentences, and 6,955 text-bound entity annotations across the training and test splits. The dataset covers 15 coarse-grained and 31 fine-grained entity types spanning diagnostically relevant clinical evidence in prion disease literature. The training split contains 247 abstracts, and the finalized test split contains 70 abstracts. As described in Section 3.3, the released test split reflects a single merged annotation layer derived from the adjudicated 70-abstract double-annotated subset. To support reproducibility, Appendix Section A, Appendix Section B, and Appendix Sections A.4–A.5 document the PubMed search query, the abstract-screening prompt, the annotation guidelines, and the model settings used in this work. The dataset annotations, annotation guidelines, and evaluation scripts are publicly available at `https://github.com/daotuanan/PrionNER/`.

Coarse-grained Type	Fine-grained Types	Definition Summary
Case Input (Willis, 2006)		
Age	Age	Age or age-at-onset expressions.
Symptom	Symptom	Symptoms, signs, and other explicit clinical manifestations.
Test_name	Imaging_test, Electrophysio_test, Blood_biomarker_test, Genetic_test, Molecular_assay, Autopsy	Diagnostic procedures, assays, and examination types.
Sequences	Imaging_sequence	Imaging acquisition or sequence terms.
Anatomic_location	Anatomic_location	Anatomical sites linked to findings or symptoms.
Findings	Imaging_finding, Autopsy_finding	Radiologic and pathological findings.
Case Diagnosis (Washington State Department of Health, 2022)		
Generic_Prion	Generic_Prion	Prion disease mentions without explicit subtype specification.
Sporadic_Prion	sCJD, sFI, VPSPr	Sporadic prion disease subtypes.
Familial_Prion	fCJD, GSS, FFI	Familial or inherited prion disease subtypes.
Acquired_Prion	vCJD, iCJD, Kuru	Acquired prion disease subtypes.
Differential_Diagnosis	Differential_Diagnosis	Non-prion diseases considered in differential diagnosis.
Clinical Course and Context (Thabane et al., 2013; Willis, 2006; Murad et al., 2023)		
Treatment	Treatment	Therapies, medications, and care strategies.
Complication	Complication	Secondary conditions or end-stage outcomes.
Time	Duration, Time_point	Temporal spans and specific time anchors.
Stats	Sensitivity, Specificity, Prevalence, Incidence	Diagnostic and epidemiological quantitative measures.

Table 2: Short summary of the PrionNER entity schema. Full type definitions and representative examples are provided in Appendix Table 8.

Split	Abstracts	Sentences	Entities
Train	247	2,156	5,149
Test	70	787	1,806
Total	317	2,943	6,955

Table 3: Corpus-level statistics of PrionNER.

Entity Type	Train n	Train %	Test n	Test %
Most Frequent				
Symptom	791	16.99	248	15.03
Generic_Prion	741	15.92	233	14.12
Anatomic_location	646	13.88	223	13.52
Imaging_test	261	5.61	75	4.55
Duration	217	4.66	49	2.97
Least Frequent Observed				
sFI	3	0.06	0	0.00
Prevalence	4	0.09	1	0.06
Incidence	6	0.13	4	0.24
Sensitivity	0	0.00	4	0.24
Specificity	0	0.00	4	0.24

Table 4: Selected most frequent and least frequent observed schema-defined fine-grained entity types in the train and test splits. The full split-wise fine-grained distribution is provided in Appendix Table 18.

4.2 Inter-Annotator Agreement

Inter-annotator agreement. We report agreement on the pre-adjudication independent annotations of the 70-abstract double-annotated subset described in Section 3.3. Agreement is assessed using three complementary symmetric measures: entity-level exact agreement F1, Jaccard similarity, and Cohen's kappa. Entity-level exact agreement F1 summarizes span-level agreement under the standard NER criterion that an entity counts as correct only when both annotators select the same text span and assign the same entity type (Tjong Kim Sang and De Meulder, 2003). Jaccard similarity measures the overlap between the two annotators' entity sets as the size of the intersection relative to the union (Jaccard, 1901). Cohen's kappa is computed on token-level BIO labels to quantify agreement while correcting for agreement expected by chance (Cohen, 1960). On the 70 compared documents, entity-level exact agreement is 81.78 F1. Jaccard similarity over the annotated entity sets is 69.17, and token-level BIO Cohen's kappa is 81.05. Together, these metrics provide complementary views of annotation consistency at both the span and token levels, and Appendix Table 17 provides the full per-label agreement breakdown.

4.3 Label Distribution and Structural Complexity

Tables 3 and 4 show that PrionNER covers a broad range of diagnostically relevant evidence while remaining strongly long-tailed. The most frequent fine-grained labels are *Symptom*, *Generic_Prion*, and *Anatomic_location*, whereas several clinically meaningful types remain rare. The adjudicated test set also retains structurally complex mentions, including 34 discontinuous entities, 80 nested pairs, and 2 overlapping pairs. Figure 1 gives representative examples of these discontinuous, nested, and overlapping annotations from the test set. Full split-wise label counts and structural statistics are provided in Appendix Section C. With the corpus characteristics established, we next describe the experimental setup used to benchmark the dataset.

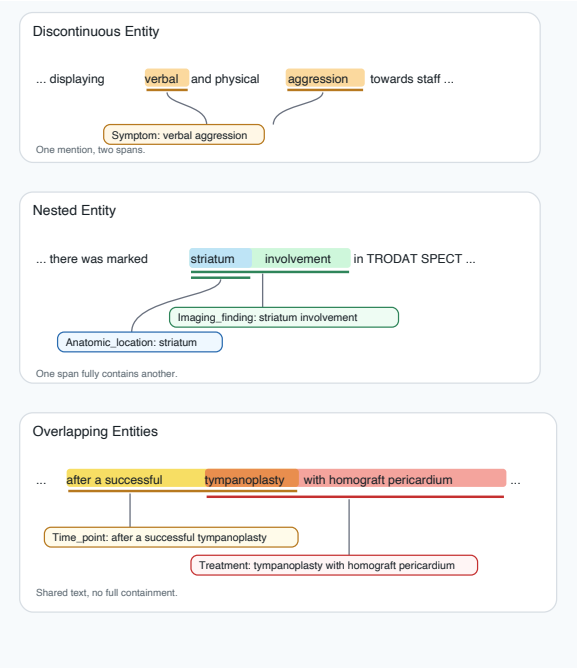


Figure 1: Representative discontinuous, nested, and overlapping entity annotations from the test set.

5 Experiments

5.1 Experimental Settings

We formulate PrionNER as a sequence-labeling task over schema-defined entity mentions. For supervised models, this sequence-labeling setup is implemented with standard BIO tagging, where each entity type is paired with `B-` and `I-` prefixes together with the outside label `O` (e.g., `B-Symptom`, `I-Symptom`, `O`). Models are trained on the training split and evaluated on the test set, with corpus statistics summarized in Table 3. Because the corpus is relatively small and highly long-tailed, we do not introduce a separate development split, as doing so would further reduce training coverage for already sparse labels. Instead, we use fixed model configurations applied uniformly within each model family, and no hyperparameters are selected based on test-set performance. We evaluate the corpus under two structural settings. In the `flat-ner` setting, both training and evaluation are restricted to the flat-compatible portion of the corpus, so only contiguous, non-overlapping gold entities are included. In the `non-flat-ner` setting, evaluation is restricted to the structurally complex portion of the corpus, so the gold annotations include only nested, discontinuous, and overlapping entities. We evaluate the dataset under two label-granularity settings. In the *fine-grained* setting, models are trained and evaluated using the 31 schema-defined entity types. In the *coarse-grained* setting, models are trained and evaluated separately using the 15 coarse-grained categories. This coarse setting is implemented as a separate task rather than by collapsing fine-grained predictions at evaluation time. We report entity-level precision, recall, and F1-score using exact span-and-label matching. Our primary metric is micro-averaged F1 on the test set, and we report it for both coarse-grained and fine-grained settings. The search query and screening prompt are documented in Appendix Section A, and model configurations are documented in Appendix Sections A.4 and A.5, so that the dataset construction and evaluation pipeline can be reproduced more directly.

5.2 Models

5.2.1 Supervised Models

We evaluate supervised baselines fine-tuned on the PrionNER training data. These include BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2021), which are widely used pretrained encoders for biomedical and clinical text mining, together with W2NER (Li et al., 2022) as an additional structured supervised baseline. We include W2NER because it is a supervised model that can also handle the `non-flat-ner` setting, allowing a supervised comparison beyond standard BIO-style flat tagging. For W2NER, we use PubMedBERT embeddings as the underlying biomedical text representation. The encoder baselines are implemented as BIO-based token classification systems and trained under the same data and model-selection protocol. The BERT baselines operate only in the standard `flat-ner` setting, whereas W2NER is additionally evaluated in the `non-flat-ner` setting. Detailed checkpoints and training hyperparameters are described in Appendix Section A.4.

5.2.2 Zero-shot Models

We additionally evaluate extraction performance using model families that do not require task-specific supervised fine-tuning on PrionNER. These models include OpenAI GPT-5.4 (OpenAI, 2026), Gemma 4 models (Team et al., 2024), the GLiNER2 variants GLiNER2-short and GLiNER2-def (Zaratiana et al., 2025), and GLiNER-BioMed (Yazdani et al., 2025). Because these systems typically return entity strings rather than reliable character offsets, their outputs require a determinis-

Model	Coarse			Fine		
	P	R	F1	P	R	F1
Supervised						
W2NER	**84.23**	79.63	**81.86**	**83.07**	78.01	**80.46**
PubMedBERT	81.88	**81.37**	81.62	78.93	**80.06**	79.49
BioBERT	80.96	81.47	81.22	78.61	78.61	78.61
ClinicalBERT	79.45	79.10	79.28	78.29	78.73	78.51
Zero-shot						
GPT-5.4	72.50	58.81	64.94	72.01	58.42	64.50
Gemma-4-31B	**78.33**	**65.61**	**71.41**	**80.14**	**59.67**	**68.41**
Gemma-4-26B	77.43	55.25	64.48	79.71	46.67	58.87
GLiNER2-def	67.62	59.41	63.25	73.29	43.10	54.28
GLiNER-BioMed	29.50	42.71	34.90	28.99	45.35	35.37
GLiNER2-short	32.63	32.67	32.65	39.51	34.79	37.00

Table 5: `flat-ner` results for all supervised and zero-shot models on PrionNER under coarse-grained and fine-grained label spaces. Best scores are bolded separately within the supervised and zero-shot sections.

Model	Coarse			Fine		
	P	R	F1	P	R	F1
Supervised / Structured						
W2NER	**12.40**	**14.81**	**13.48**	**12.74**	**14.81**	**13.70**
Zero-shot						
GPT-5.4	4.19	9.63	5.84	4.19	**9.63**	5.84
Gemma-4-31B	4.82	**12.59**	6.97	**4.96**	**9.63**	**6.55**
Gemma-4-26B	**5.76**	11.85	**7.75**	0.56	0.74	0.63
GLiNER2-def	3.53	4.44	3.93	2.78	0.74	1.17
GLiNER-BioMed	0.00	0.00	0.00	0.00	0.00	0.00
GLiNER2-short	2.07	2.22	2.14	2.90	2.96	2.93

Table 6: `non-flat-ner` results on PrionNER under coarse-grained and fine-grained label spaces. This setting includes only nested, discontinuous, and overlapping entities. Best scores are bolded separately within the supervised/structured and zero-shot sections.

tic span-alignment step before exact-match evaluation. For GLiNER2, we consider two input modes: GLiNER2-short, which provides only entity type names, and GLiNER2-def, which provides entity type names together with their definitions. GLiNER-BioMed is evaluated only in the *short* setting. In our current GLiNER-based pipeline, non-flat behavior can be represented for nested and overlapping entities through multiple contiguous spans, but the model does not support explicit discontinuous multi-span entity objects. We also explored additional small local biomedical LLMs, including `Llama3-OpenBioLLM-8B` and `BioMistral-7B`, but found them unreliable for strict exact-span extraction in this pipeline. This setting allows us to assess how well general or biomedical zero-shot extractors transfer to rare-disease clinical literature without supervised adaptation. Detailed model checkpoints, prompting, and inference settings are described in Appendix Section A.5.

5.3 Main Results

Tables 5 and 6 report the main results under coarse-grained and fine-grained label spaces in the `flat-ner` and `non-flat-ner` settings. All models are evaluated in `flat-ner`, whereas `non-flat-ner` is reported only for W2NER and the zero-shot models because the BERT baselines operate only in the standard flat setting. This design keeps discontinuous and other structurally complex gold annotations in the benchmark rather than discarding them.

In the `flat-ner` setting, W2NER is the strongest supervised model in both label spaces, reaching 81.86 F1 in the coarse-grained setting and 80.46 F1 in the fine-grained setting. Among the zero-shot systems, Gemma-4-31B performs best, with 71.41 coarse-grained F1 and 68.41 fine-grained F1.

A notable pattern in `flat-ner` is the relatively high precision of the strongest zero-shot models under the span-recovery pipeline described in Appendix Section A.5. Gemma-4-31B reaches 78.33 precision in the coarse-grained setting and 80.14 precision in the fine-grained setting, suggesting that the schema is semantically coherent enough for strong zero-shot label assignment from definitions and task instructions alone, even though recall remains well below the supervised baselines. Appendix Sections A.5 and D.3 further suggest that the strongest zero-shot model is limited more by omission than by label precision: Gemma-4-31B remains close to PubMedBERT in entity-only precision while trailing much more clearly in recall.

In the `non-flat-ner` setting, performance drops sharply for all models, but W2NER remains clearly stronger than the zero-shot systems, reaching 13.48 coarse-grained F1 with 12.40 precision and 14.81 recall, and 13.70 fine-grained F1 with 12.74 precision and 14.81 recall. Among the zero-shot systems, the best coarse-grained `non-flat-ner` score is 7.75 for Gemma-4-26B, while the best fine-grained score is 6.55 for Gemma-4-31B. GPT-5.4 is second among these models in the fine-grained `non-flat-ner` setting with 5.84 F1 and remains competitive in the coarse-grained setting, but it does not lead there. Within the GLiNER family, coarse-grained `non-flat-ner` performance improves slightly relative to fine-grained evaluation, but the GLiNER variants remain clearly behind the

top LLMs, and GLiNER-BioMed fails to recover any non-flat entities in this setting.

Overall, fine-grained prediction is generally harder than coarse-grained prediction, confirming that finer label distinctions remain substantially harder, especially once structurally complex mentions are retained in evaluation. Appendix Figure 2 further visualizes the fine-grained precision–recall trade-offs across the evaluated models.

5.4 Per-Type Analysis

Appendix Figure 3 summarizes the per-type fine-grained F1 patterns for the main supervised and zero-shot models, and Appendix Table 16 gives the full scores. Performance is uneven across the schema. Supervised models dominate many frequent labels such as *Anatomic_location*, *Generic_Prion*, *Imaging_test*, *Symptom*, and *sCJD*, which helps explain their strong aggregate performance. Zero-shot models remain competitive on several semantically clearer or clinically distinctive labels, with GPT-5.4 or Gemma 4 achieving the best score for types such as *FFI*, *Electrophysio_test*, *fCJD*, *Genetic_test*, *Molecular_assay*, and *Differential_Diagnosis*. GLiNER2-def is a notable exception within the GLiNER family, reaching 1.00 F1 on *Sensitivity*, but sparse or boundary-sensitive labels such as *Time_point*, *iCJD*, *Imaging_finding*, and especially *Prevalence* remain difficult for nearly all systems. These extreme per-type values should be interpreted cautiously for very rare labels, where isolated perfect or zero scores may reflect only a handful of test mentions rather than stable schema-wide behavior.

6 Discussions

The main results and per-type patterns motivate a broader discussion of why the benchmark remains difficult and how well it aligns with prion-disease clinical evidence.

6.1 Remaining Challenges

First, the entity distribution is long-tailed. As shown in Table 4, a small number of frequent labels account for a large share of mentions: *Symptom*, *Generic_Prion*, and *Anatomic_location* account for 791, 741, and 646 mentions in the training split and 248, 233, and 223 in the test split, respectively. At the same time, many clinically meaningful categories remain sparse, including *Prevalence* (4 in the training split, 1 in the test split) and *Sensitivity* and *Specificity* (0 in the training split and 4 each in the test split). This skew is not merely a corpus artifact but reflects the structure of real diagnostic reporting, where a few common evidence types dominate while rarer findings and subtypes remain clinically important. Difficulty also reflects lexical diversity rather than frequency alone: even common labels such as *Symptom*, *Anatomic_location*, and *Duration* cover hundreds of distinct normalized surface forms in the training split. There is also substantial train–test surface-form mismatch, so models must generalize beyond memorized mention dictionaries even for relatively common categories.

Second, the schema requires distinctions among diagnostically adjacent categories, including prion subtypes, test names, findings, temporal expressions, and differential diagnoses, so errors are not only boundary errors but often clinically meaningful type confusions. In practice, subtype mentions such as *fCJD*, *iCJD*, and "corneal transplant-related CJD" are sometimes predicted as *Generic_Prion*. Others arise when late-stage downstream consequences such as *dysphagia* or *death* appear lexically symptom-like even though they are better labeled as *Complication* in context. The imaging cluster is also challenging, with *Imaging_test*, *Imaging_sequence*, and *Imaging_finding* often requiring fine-grained distinction among modality names, acquisition terms, and radiologic abnormalities. Appendix Section D.3 provides a more detailed entity-only confusion analysis with illustrative examples.

Third, the corpus contains discontinuous, nested, and overlapping mentions that are difficult for standard extraction pipelines. This difficulty is partly a consequence of dataset design itself: PrionNER preserves clinically adjacent labels and non-flat mention structures because simplifying them away would reduce the realism of the target task. At the structural level, the training split contains 97 discontinuous entities, 235 nested pairs, and 5 overlapping pairs, while the test split contains 34 discontinuous entities, 80 nested pairs, and 2 overlapping pairs. In practice, this means that most structural difficulty comes from nesting rather than true overlap. Appendix Tables 10–13 provide the full structural breakdown, and Figure 1 illustrates representative examples from the test set.

6.2 Clinical Alignment

PrionNER captures key prion-disease characteristics, including symptoms, tests, and disease pro-

gression described in the academic literature and clinical guideline documents, rather than arbitrary lexical patterns (Washington State Department of Health, 2022; Centers for Disease Control and Prevention, 2026). CJD-related terminology is strongly represented in the labeled terms, with *CJD* appearing 576 times, which is consistent with the prominence of Creutzfeldt-Jakob disease in scientific reports. The `Symptom` entity (1,279 mentions) aligns closely with current medical guidance and captures hallmark clinical features such as *dementia* (79), *myoclonus* (79), and *ataxia* (47). The dataset also preserves rarer but clinically important prion-disease subtypes and variants such as FFI (85), GSS (63), and the Heidenhain variant (4). Finally, the `Duration` entity (290 mentions) captures meaningful disease-course information, with frequent expressions such as *within a year* (7) and *12 months* (7), reflecting the typical one-year survival duration described in the clinical literature. Taken together, these patterns show that PrionNER aligns well with expert knowledge and published guidelines. This makes the dataset useful not only for benchmarking NER systems but also for downstream knowledge extraction and future diagnostic-support settings.

7 Conclusions

In this paper, we introduce PrionNER, a manually annotated named entity recognition dataset for prion disease biomedical literature derived from 317 PubMed abstracts. We present a clinically grounded schema that captures fine-grained diagnostic evidence and non-flat entity structure, together with a benchmark spanning supervised and zero-shot extraction settings. PrionNER provides a reliable annotation resource, with 81.78 entity-level exact agreement F1 on the pre-adjudication double-annotated test split. Our experiments show that W2NER provides the strongest supervised results, PubMedBERT is the strongest BERT baseline, and Gemma-4-31B is the strongest zero-shot model, while the remaining performance gaps confirm that PrionNER is a useful but challenging benchmark for prion-disease information extraction. More broadly, the benchmark surfaces challenges that are likely to recur in other rare-disease settings, including sparse but clinically important labels, fine-grained diagnostic distinctions, and nested or discontinuous spans. We hope PrionNER will support future research on rare-disease information extraction, structured knowledge construction, and clinically oriented biomedical NLP, including relation extraction, document-level evidence consolidation, and normalization or retrieval settings that connect extracted mentions to structured rare-disease knowledge resources.

Limitations

PrionNER has several limitations. First, it is built from 317 PubMed abstracts, which limits corpus size and coverage. Second, its focus on prion diseases may reduce generalizability to other disorders. Third, the data come from biomedical abstracts rather than clinical notes, so the language is more summarized and structured than real-world records. Future work should expand the corpus, improve coverage of sparse labels, and incorporate additional clinical text sources.

Ethics Statement

This study uses publicly available PubMed abstracts and does not involve human subjects or protected health information. No identifiable personal data were collected or processed, so institutional review board approval was not required.

Acknowledgement

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner Jr, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):161.

Rebeca Benavente and Rodrigo Morales. 2024. Therapeutic perspectives for prion diseases in humans and animals. *PLoS pathogens*, 20(12):e1012676.

Centers for Disease Control and Prevention. 2026. About prion diseases. https://www.cdc.gov/prions/about/index.html. Accessed: 2026-04-15.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, pages 70–75.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. Meddec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455.

Nils Gehlenborg, Daehee Hwang, Inyoul Y Lee, Hyuntae Yoo, David Baxter, Brianne Petritis, Rose Pitstick, Bruz Marzolf, Stephen J DeArmond, George A Carlson, et al. 2009. The prion disease database: a comprehensive transcriptome resource for systems biology research in prion diseases. *Database*, 2009:bap011.

Michael D. Geschwind. 2015. Prion diseases. *CONTINUUM: Lifelong Learning in Neurology*, 21(6, Neuroinfectious Disease):1612–1638.

Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. 2015. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015:bav005.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Mika Inada Shimamura and Katsuya Satoh. 2025. Challenges and revisions in diagnostic criteria: Advancing early detection of prion diseases. *International Journal of Molecular Sciences*, 26(5):2037.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. 2022. The raredis corpus: a corpus annotated with rare diseases, their signs and symptoms. *Journal of biomedical informatics*, 125:103961.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

M Hassan Murad, Lifeng Lin, Haitao Chu, Bashar Hasan, Reem A Alsibai, Alzhraa S Abbas, Reem A Mustafa, and Zhen Wang. 2023. The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy. *Cmaj*, 195(27):E925–E931.

OpenAI. 2026. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-04-11.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. *CLEF (working notes)*, 1179.

Stanley B Prusiner. 1998. Prions. *Proceedings of the National Academy of Sciences*, 95(23):13363–13383.

Andrea Saltelli, Ksenia Aleksankina, William Becker, Pamela Fennell, Federico Ferretti, Niels Holst, Sushan Li, and Qiongli Wu. 2019. Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental modelling & software*, 114:29–39.

Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. Identifying and extracting rare diseases and their phenotypes with large language models. *Journal of Healthcare Informatics Research*, 8(2):438–461.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

S Tenny and MR Hoffman. 2017. Prevalence.

Lehana Thabane, Lawrence Mbuagbaw, Shiyuan Zhang, Zainab Samaan, Maura Marcucci, Chenglin Ye, Marroon Thabane, Lora Giangregorio, Brittany Dennis, Daisy Kosa, et al. 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC medical research methodology*, 13(1):92.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ozlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sonia M Vallabh, Eric V Minikel, Stuart L Schreiber, and Eric S Lander. 2020. Towards a treatment for genetic prion disease: trials and biomarkers. *The Lancet Neurology*, 19(4):361–368.

Washington State Department of Health. 2022. Prion disease reporting and investigation guideline. https://doh.wa.gov/sites/default/files/2025-08/420-069-Guideline-Prion.pdf. Last revised December 2022; accessed 2026-04-15.

Marjorie Canfield Willis. 2006. *Medical terminology: The language of health care*. Lippincott Williams & Wilkins.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface. *arXiv preprint arXiv:2507.18546*.

A Data Collection and Model Setup

A.1 PubMed Search Query

The following Boolean query is used to retrieve candidate abstracts from PubMed:

```
(
  "Prion Diseases"[Title/Abstract] OR
  "Creutzfeldt-Jakob Disease"[Title/Abstract] OR
  "CJD"[Title/Abstract] OR
  "sporadic CJD"[Title/Abstract] OR
  "familial CJD"[Title/Abstract] OR
  "genetic CJD"[Title/Abstract] OR
  "variant CJD"[Title/Abstract] OR
  "iatrogenic CJD"[Title/Abstract] OR
  "Kuru"[Title/Abstract] OR
  "Gerstmann-Straussler-Scheinker"[Title/Abstract] OR
  "Fatal Familial Insomnia"[Title/Abstract] OR
  "FFI"[Title/Abstract]
)
AND
(
  diagnosis[Title/Abstract] OR
  clinical[Title/Abstract] OR
  symptoms[Title/Abstract] OR
  case[Title/Abstract] OR
  progression[Title/Abstract] OR
  treatment[Title/Abstract]
)
NOT
(
  mice[Title/Abstract] OR
  mouse[Title/Abstract] OR
  rat[Title/Abstract] OR
  animal[Title/Abstract] OR
  cell[Title/Abstract] OR
  protein[Title/Abstract] OR
  in vitro[Title/Abstract]
)
```

A.2 Prompt Used for Abstract Relevance Screening

Before applying model-based screening to the full retrieval set, two annotators manually screened approximately 500 abstracts and observed that a large fraction of the keyword-matched results were not actually suitable for corpus construction. The most common false matches were papers centered on basic science, animal or other non-human studies, non-clinical analyses such as economic or purely epidemiological reports, or papers in which prion disease was only a secondary topic. We used this pilot review to formulate an operational definition of *related* abstracts for automated screening.

The GPT-5.4 screening prompt asks the model to determine whether an abstract is relevant to *human prion diseases in a clinical context*, based only on the abstract text. Relevant abstracts are defined as those primarily focused on diagnosis, symptoms, disease progression, or treatment in human prion disease. The prompt explicitly instructed the model to reject abstracts centered on basic science, animal or non-human studies, protein mechanisms without clinical human focus, unrelated primary diseases, or non-clinical reports such as economic, policy, surveillance, or purely epidemiological analyses.

The model is required to return a strict JSON object with three fields: ‘is_relevant‘, ‘reason‘, and ‘evidence_spans‘. If relevance checking is enabled, we verify that the response is a JSON object with `is_relevant` as a Boolean value, `reason` as a non-empty string, and `evidence_spans` as a list. We then retain only evidence spans that can be matched to the source abstract after light text normalization. A simplified version of the prompt structure is shown below.

```
System: You are a biomedical abstract screening
    assistant.
Decide whether the abstract is relevant to human
     prion
diseases in a clinical context.
Return one strict JSON object only.

User:
- Use the provided relevance definition.
- Judge based only on the abstract text.
- Exclude basic science, animal, non-human,
    unrelated-topic, and non-clinical abstracts.
- Return JSON only in the form:


{
  "is_relevant": true,
  "reason": "<short reason>",
  "evidence_spans": [
    "<short exact span from abstract>"
  ]
}
```

A.3 Audit of the Relevance Filter

After the initial retrieval and preprocessing stage produced 3,138 abstracts, GPT-5.4 screened the full set with the relevance definition above and labeled 1,304 abstracts as related and 1,834 as unrelated. We then manually reviewed abstracts from the screened pool together with the earlier pilot screening results to confirm relevance and remove duplicates. In total, this yielded 1,383 manually reviewed abstracts for auditing the screening process, including 868 abstracts rated as related by GPT-5.4.

Within this manually reviewed set, 772 abstracts were judged truly related and 611 were judged truly unrelated. GPT-5.4 predicted 868 abstracts as related and 515 as unrelated. Among the 868 abstracts rated as related by GPT-5.4, 755 were truly related and 113 were actually unrelated, corresponding to a precision of 86.98 for the related class. The resulting confusion pattern was strongly asymmetric: only 17 truly related abstracts were missed, whereas 113 unrelated abstracts were incorrectly marked as related. This shows that GPT-5.4 was aggressive in calling abstracts relevant.

On this 1,383-abstract audit set, GPT-5.4 achieved 90.60 accuracy and 89.65 balanced accuracy. For the related class, precision was 86.98, recall was 97.80, and F1 was 92.07. For the unrelated class, precision was 96.70, recall was 81.51, and F1 was 88.45. These headline metrics indicate that about 90.60% of all abstracts were classified correctly. Balanced accuracy of 89.65% is a useful summary here because it averages recall on the related and unrelated classes, which is important given the model's asymmetric behavior. For the related class, GPT-5.4 was correct 86.98% of the time when it predicted *related*, and it recovered 97.80% of all truly related abstracts, indicating that it was very effective at not missing relevant papers. For the unrelated class, GPT-5.4 was correct 96.70% of the time when it predicted *not related*, but its recall for that class was lower at 81.51%, meaning that it did not label unrelated abstracts as *not related* often enough. In practical terms, GPT-5.4 functioned as a strong high-recall relevance screener: it is well suited to settings where the priority is to avoid missing relevant abstracts, but it is weaker at filtering out irrelevant abstracts cleanly. In plain terms, the model tends to over-include, which may be acceptable for screening but increases downstream manual review effort and can introduce more false positives for later NER annotation.

A.4 Supervised Model Details

The supervised BERT model aliases and checkpoints are as follows: `biobert` = `dmis-lab/biobert-base-cased-v1.2`, `clinicalbert` = `emilyalsentzer/Bio_ClinicalBERT`, and `pubmedbert` = `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`. To avoid further reducing training coverage for rare labels, we do not create a separate development split; instead, we fix the training hyperparameters in advance and apply the same settings across the BERT baselines. No hyperparameter choices are made based on test-set results. All BERT models use the same hyperparameters: `max_length` = 512, `num_train_epochs` = 5.0, `learning_rate` = 5e-05, `weight_decay` = 0.01, `train_batch_size` = 8, `eval_batch_size` = 16, and `seed` = 42. BERT flat-ner results for this 5-epoch setup are as follows. In coarse-grained evaluation, BioBERT reaches 80.96 precision, 81.47 recall, and 81.22 F1; ClinicalBERT reaches 79.45 precision, 79.10 recall, and 79.28 F1; and PubMedBERT reaches 81.88 precision, 81.37 recall, and 81.62 F1. In fine-grained evaluation, BioBERT reaches 78.61 precision, 78.61 recall, and 78.61 F1; ClinicalBERT reaches 78.29 precision, 78.73 recall, and 78.51 F1; and PubMedBERT reaches 78.93 precision, 80.06 recall, and 79.49 F1. For W2NER (Li et al., 2022), we use PubMedBERT embeddings as the encoder input representation and train for 10 epochs. We include it specifically because it remains supervised while also supporting the `non-flat-ner` setting, unlike the BIO-based BERT baselines. We additionally report aggregate W2NER runs in both `flat-ner` and `non-flat-ner` settings. In coarse-grained `flat-ner`, W2NER reaches 84.23 precision, 79.63 recall, and 81.86 F1; in fine-grained `flat-ner`, it reaches 83.07 precision, 78.01 recall, and 80.46 F1. For `non-flat-ner`, W2NER reaches 12.40 precision, 14.81 recall, and 13.48 F1 in the coarse-grained setting, and 12.74 precision, 14.81 recall, and 13.70 F1 in the fine-grained setting.

A.5 Zero-shot Prompting and Inference Details

For zero-shot evaluation, we apply the GLiNER2 variants GLiNER2-short and GLiNER2-def (Zaratiana et al., 2025) together with GLiNER-BioMed (Yazdani et al., 2025), without supervised fine-tuning on PrionNER. For both GLiNER2-short and GLiNER2-def, we use the checkpoint `fastino/gliner2-large-v1`; for GLiNER-BioMed, we use `Ihor/gliner-biomed-base-v1.0`. GLiNER2-short supplies only the entity type names, whereas GLiNER2-def supplies both the entity type names and short type definitions. GLiNER-BioMed is evaluated only with entity type names. In our current GLiNER-based pipeline, non-flat behavior can be represented for nested and overlapping entities through multiple contiguous spans, but the model does not support explicit discontinuous multi-span entity objects. We also tested smaller local biomedical LLMs, including `aaditya/Llama3-OpenBioLLM-8B` and `BioMistral/BioMistral-7B`, but excluded them from the main comparison because in our pipeline they frequently produced chat-style outputs, malformed or truncated JSON, corrupted labels, and non-literal spans, making them unreliable for strict exact-span NER extraction. The models are run at the abstract level.

For the LLM-based setting, we evaluate OpenAI GPT-5.4 (OpenAI, 2026), `google/gemma-4-31B-it`, and `google/gemma-4-26B-A4B-it` from the Gemma 4 family (Team et al., 2024) using a fixed prompt template that introduces the task, specifies the target entity schema, and requests structured outputs. Each input instance is processed at the abstract level, and the model outputs are required to follow a constrained format such as JSON with entity text.

After inference, we apply a verification and normalization pipeline before evaluation. For NER outputs, we first verify the output shape: the model response must be a JSON object containing an `entities` list, and for OpenAI/Gemma outputs the returned `text` field must match the input abstract after collapsing whitespace. We then validate entity labels against the schema, rejecting invalid `coarse_type`/`fine_type` combinations. In limited cases, we repair partially correct predictions: in the coarse-only schema setting, `fine_type` is normalized to the same value as `coarse_type`, and if a predicted `fine_type` is valid but the `coarse_type` is missing or incorrect, we infer the corresponding coarse label from the schema. Next, we verify that each predicted mention can be aligned back to the source text and attach start/end offsets with deduplication. This alignment step is necessary for the zero-shot systems because LLM- and GLiNER-style outputs typically return entity strings but not reliable character offsets, whereas exact-match NER evaluation requires position-specific spans. When the same predicted surface form occurs multiple times in the same abstract, we map that mention to all exact string-matching occurrences in order to recover candidate offsets deterministically. This expansion step should therefore be understood as an offset-recovery procedure for span-text outputs rather than an additional modeling component, although it introduces an evaluation asymmetry relative to supervised token-level baselines that predict positions directly. For OpenAI/Gemma outputs, mentions that cannot be aligned are skipped and logged with warnings.

The GPT-5.4 prompt template used in our experiments is shown below.

```
System:
You are an information extraction model for
    biomedical case reports and reviews about
    prion disease.
Your job is to label entity spans using only the
    provided schema and output one valid JSON
    object.

User: Read the schema carefully and follow it
    exactly.

Schema JSON: {entity_schema}

Rules:
1. Extract only explicit spans from the text.
2. Use exact labels from the schema.
3. Use exact text spans from the input.
4. Do not output start or end offsets.
5. Do not output comments, markdown, or extra
    keys.
6. If uncertain, omit the span instead of
    guessing.

Input text: {text}

Output exactly one JSON object with this
    structure:
{
  "text": "<original input text>",
  "entities": [
    {
      "mention": "<exact span>",
      "coarse_type": "<schema coarse type>",
      "fine_type": "<schema fine type>",
      "normalized": "<normalized form or same as
      mention>"
    }
  ]
}
```

B Annotation Guidelines and Schema

B.1 Guideline Principles

Global rules: annotate only explicit mentions; prefer the most specific label available; keep spans minimal but complete; annotate both long form and abbreviation when both are explicit; treat molecular subtype strings such as *MM1*, *MV2*, and *VV2* as `sCJD` mentions, with finer subtype interpretation handled in normalization; treat exposure or treatment phrases such as *corneal transplantation*, *cadaveric-derived hormone treatment* (e.g., growth hormone), and *cadaveric dura mater graft* as `Treatment` mentions rather than disease mentions; include generic heads such as *disease*, *syndrome*, and *symptom* when they are part of the explicit annotated mention; and for statistical expressions annotate only the value span (e.g., *100%*, *0.06%*) rather than surrounding cue phrases.

Do not assign labels by keyword alone. A span is a disease mention only if it names a disease entity in context, and a span is a symptom mention only if it expresses a clinical manifestation in context. For example, *disease duration* is not a disease label, and *symptom onset* is not a `Symptom` label.

B.2 Train Annotation Consistency and Guideline Drift

The training annotations were produced over multiple rounds by the two annotators and then consolidated through adjudication into a single final corpus version. Because the train split was not preserved as two separately frozen annotator-specific versions, we do not report a formal per-annotator label-distribution comparison here. Instead, Table 7 summarizes the main recurrent sources of early annotation-style drift that were resolved during guideline refinement and then applied consistently in the final schema. Most of these changes affected boundary selection or distinctions between nearby label types rather than the overall entity inventory.

No.	First Guideline / Early Practice	Final Guideline
1	*brain* was not labeled as `Anatomic_location`.	*brain* is labeled as `Anatomic_location`.
2	*died* and *death* were not labeled as `Complication`.	*died* and *death* are labeled as `Complication`.
3	*FLAIR*, *T1-weighted*, *T2-weighted*, *ADC*, and *DWI* were labeled as `Imaging_test`.	These sequence terms are labeled as `Imaging_sequence` rather than `Imaging_test`.
4	Time expressions such as *in year 2013*, *in 2013*, and *12 months after onset of symptoms* were not consistently separated into anchor vs. duration components.	`Time_point` captures anchor expressions such as *2013* and *onset of symptoms*, while `Duration` captures spans such as *12 months*.
5	`Autopsy` was restricted to explicit procedure mentions such as *autopsy* and *autopsy examination*.	`Autopsy` was expanded to include additional procedure-reporting forms such as *pathologically*, *autopsy*, and *neuropathological examination*.
6	Statistical entities were often annotated as full cue phrases, e.g., *sensitivity of 61%* or *prevalence of 1/1 million people per year*.	Only the value span is annotated: e.g., *61%* as `Sensitivity` and *1/1 million people per year* as `Prevalence`.
7	*Heidenhain variant* was not labeled as `sCJD`.	*Heidenhain variant* is labeled as `sCJD`.

Table 7: Examples of guideline drift observed during training annotation and the final adjudicated guideline used in PrionNER.

B.3 Full Entity Schema

Table 8: Full coarse-grained and fine-grained entity schema of PrionNER, including definitions and representative examples.

Coarse-grained Type	Fine-grained Type	Definitions
Case Input (Geschwind, 2015)		
Age	Age	The age of a patient at disease onset. Examples: age 62; 45-year-old; at 58 years
Symptom	Symptom	Any subjective experience reported by the patient or objective observation made by a clinician. *Examples:* memory loss; fatigue; hyperreflexia; Babinski sign; rigidity; startle response
Test_name	Imaging_test	MRI; CT scan; PET scan
	Electrophysio_test	nerve conduction study, Polysomnography
	Blood_biomarker_test	CSF 14-3-3 assay; tau assay; complete blood count
	Genetic_test	PRNP genetic testing; mutation analysis; gene panel testing
	Molecular_assay	RT-QuIC; PCR; Western blot
	Autopsy	brain autopsy; neuropathological examination; postmortem analysis
Sequences	Imaging_sequence	DWI; FLAIR; diffusion-weighted imaging; ADC maps
Anatomic_location	Anatomic_location	A specific anatomical structure where a clinical finding or pathological change is observed. *Examples:* cortex; basal ganglia; thalamus; caudate nucleus; cerebral cortex
Findings	Imaging_finding	pulvinar sign; restricted diffusion; cortical ribboning
	Autopsy_finding	spongiform change; neuronal loss; astrogliosis
Case Diagnosis (Centers for Disease Control and Prevention, 2026)		

Continued on next page

Table 8: Full coarse-grained and fine-grained entity schema of PrionNER, including definitions and representative examples (continued).

Coarse-grained Type	Fine-grained Type	Definitions
Generic_Prion	Generic_Prion	Prion disease mentions without specific subtype classification. *Examples:* prion disease; CJD, bovine spongiform encephalopathy, BSE, mad cow disease, Transmissible Spongiform Encephalopathy, TSE
Sporadic_Prion	sCJD	Sporadic Creutzfeldt-Jakob Disease
	sFI	Sporadic Fatal Insomnia
	VPSPr	Variably Protease-Sensitive Prionopathy
Familial_Prion	fCJD	Familial Creutzfeldt-Jakob Disease
	GSS	Gerstmann-Sträussler-Scheinker Syndrome
	FFI	Fatal Familial Insomnia
Acquired_Prion	vCJD	Variant Creutzfeldt-Jakob Disease
	iCJD	Iatrogenic Creutzfeldt-Jakob Disease
	Kuru	Kuru
Differential_Diagnosis	Differential_Diagnosis	Non-prion diseases used for differential diagnosis (no fine-grained subtypes in this schema). *Examples:* Alzheimer's disease; autoimmune encephalitis; viral encephalitis; Parkinson's disease
Clinical Course and Context		
Treatment	Treatment	A therapeutic intervention, medication used to manage a disease or its symptoms. Examples: supportive care; quinacrine; doxycycline; symptomatic treatment (Benavente and Morales, 2024)
Complication	Complication	A secondary medical condition that arises as a consequence of disease progression. Examples: pneumonia; aspiration pneumonia; respiratory failure (Willis, 2006)
Time	Duration	A span or length of time over which a clinical event, symptom, or disease progression occurs. *Examples:* within 3 months; over 2 years; rapidly progressive over weeks
	Time_point	A specific point in time associated with clinical symptoms or a specific event. *Examples:* at onset; at age 62; in 2021
Stats	Sensitivity	The true positive rate for a diagnostic test (Saltelli et al., 2019). *Examples:* sensitivity of 85%
	Specificity	The true negative rate for a diagnostic test (Saltelli et al., 2019). *Examples:* specificity of 92%
	Prevalence	Disease frequency within a population (Tenny and Hoffman, 2017). *Examples:* 1 per million; prevalence of 0.5%
	Incidence	The rate at which new cases of a disease occur in a population during a specified time period. *Examples:* annual incidence of 2 per million

B.4 Compact Annotation Guideline

Table 9: Compact PrionNER annotation guideline.

Label	Use For	Annotate	Do Not Annotate
`Generic_Prion`	Prion disease mentions without explicit subtype classification	*prion disease*, *prion diseases*, *Creutzfeldt-Jakob disease*, *CJD*, *bovine spongiform encephalopathy*, *BSE*, *mad cow disease*	Specific subtype mentions such as *sporadic CJD* or *variant CJD*; protein mentions such as *prion protein*, *PrP*, or PrP^{Sc}; pathological finding uses such as *spongiform encephalopathy* when used as an autopsy finding
`sCJD`	Sporadic Creutzfeldt-Jakob disease	*sporadic Creutzfeldt-Jakob disease*, *sporadic CJD*, *sCJD*, *MM1*, *MV1*, *VV2* when clearly used as sCJD subtypes	Generic *CJD* without a sporadic cue
`sFI`	Sporadic fatal insomnia	*sporadic fatal insomnia*, *sFI*	*fatal familial insomnia*, *FFI*, *FFI-1*, *FFI-2*

Label	Use For	Annotate	Do Not Annotate
VPSPr	Variably protease-sensitive prionopathy	*variably protease-sensitive prionopathy*, *VPSPr*	*PrP*, *PrP^{Sc}*, *PrP-res*, or other protein-level mentions
fCJD	Familial or hereditary CJD	*familial CJD*, *familial Creutzfeldt-Jakob disease*, *hereditary CJD*, *genetic CJD*, *fCJD*	Mutation or genotype mentions alone
GSS	Gerstmann-Sträussler-Scheinker syndrome	*Gerstmann-Sträussler-Scheinker syndrome*, *GSS*, *GSS102*, *GSS105*	Mutation names alone; pathological descriptors such as *kuru plaques*
FFI	Fatal familial insomnia	*fatal familial insomnia*, *FFI*, *FFI-1*, *FFI-2*	Isolated *insomnia*; use Symptom instead
vCJD	Variant CJD	*variant Creutzfeldt-Jakob disease*, *variant CJD*, *vCJD*, *new variant CJD*, *nvCJD*	Generic uses of *variant* without clear disease reference
iCJD	Iatrogenic CJD	*iatrogenic Creutzfeldt-Jakob disease*, *iatrogenic CJD*, *growth hormone-associated CJD*, *dural graft associated CJD*, *iCJD*, *dCJD*	*iatrogenic* alone when the CJD referent is unclear
Kuru	Kuru disease mention	*Kuru*, *kuru*	*kuru plaques*; use Autopsy_finding instead
Differential_ Diagnosis	Non-prion diseases considered as alternatives or comparators	*Alzheimer disease*, *autoimmune encephalitis*, *Huntington disease*	Prion diseases; symptom phrases unless they are clearly used as alternative diagnoses
Symptom	Symptoms, signs, and clinical manifestations	*dementia*, *myoclonus*, *cerebellar ataxia*, *psychotic symptoms*, *hyperreflexia*	Disease names; imaging findings; autopsy findings
Imaging_ test	Imaging modalities or procedures	*MRI*, *CT*, *PET*, *SPECT*	Sequences such as *FLAIR*, *DWI*; findings such as *pulvinar sign*
Electrophysio_ test	Electrophysiology procedures	*EEG*, *electroencephalogram*, *polysomnography*	Electrophysiologic findings or waveform targets
Blood_ biomarker_ test	Biomarker assays and specimen-as-test shorthand	*CSF*, *tau assay*, *blood tests*	Molecular methods such as *PCR*, *RT-QuIC*, or *Western blot*
Genetic_ test	Genetic testing procedures	*genetic testing*, *Prion gene analysis*, *genetic analysis*	Gene names and mutation names alone
Molecular_ assay	Molecular or biochemical assays	*RT-QuIC*, *PCR*, *Western blot*, *molecular analysis*	Biopsy or autopsy procedures; gene targets without a testing method
Autopsy	Postmortem or tissue examination procedures	*brain autopsy*, *postmortem examination*, *brain biopsy*	Pathological findings themselves; anatomy alone
Imaging_ sequence	Acquisition or sequence terms	*DWI*, *FLAIR*, *T2 weighted*, *ADC maps*	Imaging modality names; imaging abnormalities
Imaging_ finding	Radiologic abnormalities or named signs	*pulvinar sign*, *restricted diffusion*, *cortical ribboning*	Test modality alone; anatomy alone unless separately annotated
Autopsy_ finding	Pathological or postmortem findings	*spongiform change*, *neuronal loss*, *gliosis*, *florid plaques*, *kuru plaques*, *spongiform encephalopathy* when used as a pathological finding	Procedure terms; disease names
Anatomic_ location	Body locations linked to findings or symptoms	*thalamus*, *basal ganglia*, *caudate nucleus*, *cortex*, *deep white matter*	Non-anatomical descriptors; whole finding phrase when only one part is anatomy
Age	Age or age-at-onset expression	*58-year-old*, *age 62*, *young adults*	Durations; calendar dates

Label	Use For	Annotate	Do Not Annotate
`Treatment`	Therapies, medications, or care strategies	*palliative care*, *quinacrine*, *doxycycline*, *antipsychotics*; annotate the full treatment mention when expressed as a therapy or care strategy	Diagnostic tests; non-intervention goals unless expressed as actual therapy
`Complication`	Secondary adverse conditions or end-stage outcomes	*pneumonia*, *respiratory failure*, *SIADH*, *death*, *died*	The primary prion disease itself
`Duration`	Time span or length	*within 3 months*, *over 2 years*, *13 months*	Calendar dates; age expressions
`Time_point`	Specific time anchor	*at onset*, *on admission*, *1996*, *January 2012*	Durations; exclude discourse connectives such as *in*, *after* when they are not themselves the time anchor
`Sensitivity`	Diagnostic sensitivity expression	*100%*, *91%* when they are explicitly the sensitivity value, e.g., annotate only *100%* in *sensitivity of 100%*	The cue word alone, e.g., *sensitivity*, when no value is included; unrelated percentages
`Specificity`	Diagnostic specificity expression	*92%*, *95%* when they are explicitly the specificity value	The cue word alone, e.g., *specificity*, when no value is included; unrelated percentages
`Prevalence`	Disease frequency expression	*1–2 people per million annually*, *0.5%* when it is explicitly the prevalence value	The cue word alone, e.g., *prevalence*, when no value is included; sample size counts
`Incidence`	New-case rate expression	*0.06%*, *2 per million annually*, *0.37 cases/million*, *1 in 1 000 000* when they are explicitly the incidence value	The cue word alone, e.g., *incidence*, when no value is included; sample size counts; surrounding time or cue phrases such as *in 2023*, *million per year*, or *annual incidence of*

C Extended Dataset Statistics

C.1 Structural Annotation Statistics

The train and test splits both contain non-trivial structural complexity at the span level (Table 10). Discontinuous entities remain relatively uncommon in absolute terms, but they appear in both splits and are distributed across clinically salient labels rather than being confined to a single category. In both train and test, *Symptom* is the most common discontinuous label, followed by *Anatomic_location*; train also shows notable discontinuous cases for *Imaging_finding*, while test includes comparatively more *Blood_biomarker_test* discontinuities (Table 11).

Nested structures are much more frequent than overlapping ones, and they are concentrated in recurring clinically meaningful label pairs (Tables 12 and 13). In the training split, the most common nested envelope patterns are *Symptom* within *Symptom*, *Duration* with *Time_point*, and *Anatomic_location* paired with *Symptom*; in the test split, the dominant patterns shift toward *Anatomic_location* with *Symptom* or *Imaging_finding*, with an additional cluster of *Treatment–iCJD* nesting. By contrast, overlapping envelope pairs are rare in both splits, suggesting that the benchmark's structural difficulty is driven primarily by discontinuity and especially nesting rather than widespread partial-overlap phenomena.

Split	Docs	Entities	Disc.	Nested Env.	Overlap Env.
Train	247	5,149	97	235	5
Test	70	1,806	34	80	2

Table 10: Span-structure summary for the train and test splits. Discontinuous entities are text-bound entities with more than one atomic span. Nested and overlapping pair counts are reported at the entity-envelope level.

Label	Train	Test
Symptom	36	10
Anatomic_location	15	9
Imaging_finding	14	3
Blood_biomarker_test	4	5
Differential_Diagnosis	4	2
Treatment	4	1
Duration	3	2
Imaging_sequence	4	0
Imaging_test	4	0
Autopsy_finding	3	1
sCJD	3	0
vCJD	1	1
fCJD	1	0
iCJD	1	0

Table 11: Discontinuous entities by label in the train and test splits. The train split contains discontinuous entities in 54 documents, and the test split contains discontinuous entities in 20 documents.

Label A	Label B	Train	Test
Symptom	Symptom	42	9
Anatomic_ location	Symptom	28	13
Duration	Time_ point	30	2
Anatomic_ location	Anatomic_ location	18	10
Anatomic_ location	Imaging_ finding	14	11
Anatomic_ location	Autopsy_ finding	16	4
Imaging_ finding	Imaging_ sequence	9	2
Anatomic_ location	Differential_ Diagnosis	8	0
Blood_ biomarker_ test	Blood_ biomarker_ test	4	4
Treatment	iCJD	1	6
Imaging_ sequence	Imaging_ sequence	6	0
Differential_ Diagnosis	Differential_ Diagnosis	4	2
Generic_ Prion	Time_ point	4	2
Duration	Duration	2	3
Imaging_ finding	Imaging_ finding	4	0
Autopsy_ finding	Kuru	3	1
Anatomic_ location	Treatment	3	0
Imaging_ sequence	Imaging_ test	3	0
Differential_ Diagnosis	Symptom	2	1
Blood_ biomarker_ test	Differential_ Diagnosis	0	3

Table 12: Nested entity-envelope pairs by label pair in the train and test splits.

Label A	Label B	Train	Test
Anatomic_ location	Anatomic_ location	1	0
Autopsy_ finding	Generic_ Prion	1	0
FFI	sFI	1	0
Imaging_ test	Imaging_ test	1	0
Symptom	Symptom	1	0
Blood_ biomarker_ test	Differential_ Diagnosis	0	1
Time_ point	Treatment	0	1

Table 13: Overlapping entity-envelope pairs by label pair in the train and test splits.

Split	Abstracts	Entity Types	Mentions	Unique Surface Forms
Train	247	28	4,655	1,889
Test	70	29	1,650	729

Table 14: Split-level summary of schema-defined non-meta entity annotations and unique normalized surface forms.

Entity Type	Train Unique	Test Unique	Train Mentions	Test Mentions
Symptom	424	148	791	248
Anatomic_location	212	89	646	223
Duration	163	41	217	49
Time_point	127	39	152	47
Treatment	102	59	126	88
Autopsy_finding	110	32	215	42
Differential_ Diagnosis	113	27	188	36
Age	108	25	137	29
Imaging_finding	92	38	153	71
Imaging_test	72	22	261	75
Generic_Prion	61	27	741	233
Autopsy	54	25	199	64
Imaging_sequence	38	27	117	77
Electrophysio_test	42	10	173	38
Molecular_assay	20	29	23	45
Blood_ biomarker_ test	23	14	64	39
sCJD	18	16	86	80
iCJD	18	10	32	16
Complication	18	6	78	19
GSS	14	8	32	16

Table 15: Train/test comparison of unique normalized surface forms and non-meta mention counts for the most frequent entity types.

C.2 Split-Level Entity Summary

C.3 Entity Surface-Form Dictionary Summary

Unique surface forms are counted per entity type after lowercasing, trimming edge whitespace, collapsing internal whitespace, and joining discontinuous spans with spaces.

D Extended Results

D.1 Overall Performance

Figure 2 complements the main fine-grained results by showing the precision–recall trade-offs across the evaluated models. The supervised BERT baselines cluster in the strongest region overall, while the zero-shot models show a wider spread and sharper trade-offs; among them, Gemma-4-31B occupies the strongest overall position in the fine-grained setting.

D.2 Per-Type Fine-grained Results

Figure 3 and Table 16 show that fine-grained performance is highly uneven across the schema. The heatmap makes clear that there is no single difficulty level for PrionNER: some labels are solved reliably by most strong systems, while others re-

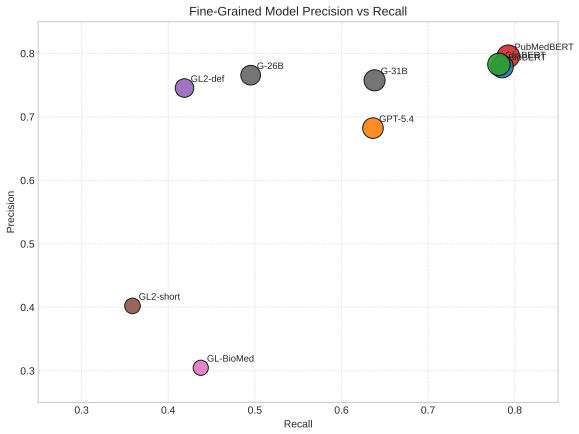


Figure 2: Precision–recall trade-offs for fine-grained entity extraction across the evaluated models.

main unstable even for the best models. This heterogeneity tracks both frequency and semantic specificity. Common and lexically anchored categories such as *Generic_Prion*, *sCJD*, *Imaging_test*, *Symptom*, *Age*, and *Anatomic_location* are comparatively robust, whereas sparse or context-dependent labels such as *Prevalence*, *Specificity*, *Sensitivity*, *Time_point*, *Molecular_assay*, and subtype distinctions such as *iCJD* and *fCJD* remain much harder.

The supervised encoders occupy the strongest region overall and are more consistent across label families. PubMedBERT in particular is among the top performers on many of the clinically common labels, reaching 91.77 on *Generic_Prion*, 90.79 on *Imaging_test*, 89.74 on *sCJD*, 88.89 on *Age*, 84.95 on *Symptom*, and 81.48 on *Blood_biomarker_test*. BioBERT and ClinicalBERT show a similar profile and outperform PubMedBERT on a few categories, such as *Anatomic_location*, *vCJD*, *Complication*, *Autopsy*, *Autopsy_finding*, *Incidence*, and *iCJD*. Taken together, the supervised models suggest that once training data are available, the main gains come from broad coverage across the entire schema rather than isolated wins on a handful of labels.

The zero-shot models show a much sharper specialization pattern. Gemma-4-31B is the strongest zero-shot model overall, but its strengths are concentrated in a narrower subset of distinctive labels, including *FFI* (100.00), *fCJD* (77.78), *Genetic_test* (71.43), *Molecular_assay* (65.79), *Duration* (65.98), and *Differential_Diagnosis* (61.29). GPT-5.4 is competitive on several labels and is best on *Electrophysio_test* (89.19) and *Time_point* (40.78), while GLiNER2 occasionally produces isolated best scores on very sparse categories such as *Sensitivity*, *Specificity*, and *Duration*. However, these zero-shot wins are scattered and do not translate into the same schema-wide stability seen in the supervised baselines.

Two further patterns are worth noting. First, the imaging and diagnostic-context families remain difficult even when overall F1 is moderate: *Imaging_finding*, *Imaging_sequence*, *Blood_biomarker_test*, *Autopsy_finding*, and *Time_point* all show larger cross-model variation than canonical disease-name labels. Second, the most extreme values should be interpreted with caution for very rare labels. For example, *Prevalence* remains at 0.00 for all models, and isolated perfect or near-perfect results on labels such as *Sensitivity* or *FFI* reflect very small test counts rather than uniformly solved clinical reasoning. Overall, the per-type view reinforces the main conclusion of the paper: PrionNER rewards models that combine strong lexical grounding on common biomedical entities with fine-grained contextual discrimination on rarer, semantically adjacent categories.

D.3 Entity-only Confusion Analysis

Figures 4a and 4b provide a complementary entity-only view of fine-grained errors for the two strongest models. Under this evaluation, PubMedBERT is clearly stronger overall, reaching 0.794 flat entity-level F1 compared with 0.684 for Gemma-4-31B. The main difference is recall: PubMedBERT reaches 0.793 recall versus 0.597 for Gemma-4-31B, while their precision remains very similar at 0.796 and 0.801, respectively. This pattern shows that Gemma-4-31B is comparatively conservative, preserving Gemma-level precision but missing many more entities.

In these row-normalized, entity-label-only matrices, PubMedBERT has the cleaner diagonal overall. More labels retain darker diagonal mass with less off-diagonal spill, indicating more consistent label assignment across the schema. Gemma-4-31B looks reasonable on a few canonical labels, but its diagonal is visibly weaker for many harder clinical categories, including *Autopsy*, *Autopsy_finding*, *Blood_biomarker_test*, *Time_point*, and *Treatment*.

Several confusion patterns are shared by both models. *Prevalence* and *Incidence* remain difficult to separate, the imaging cluster shows substantial overlap among *Imaging_finding*, *Imaging_sequence*, and *Imaging_test*, and the prion sub-

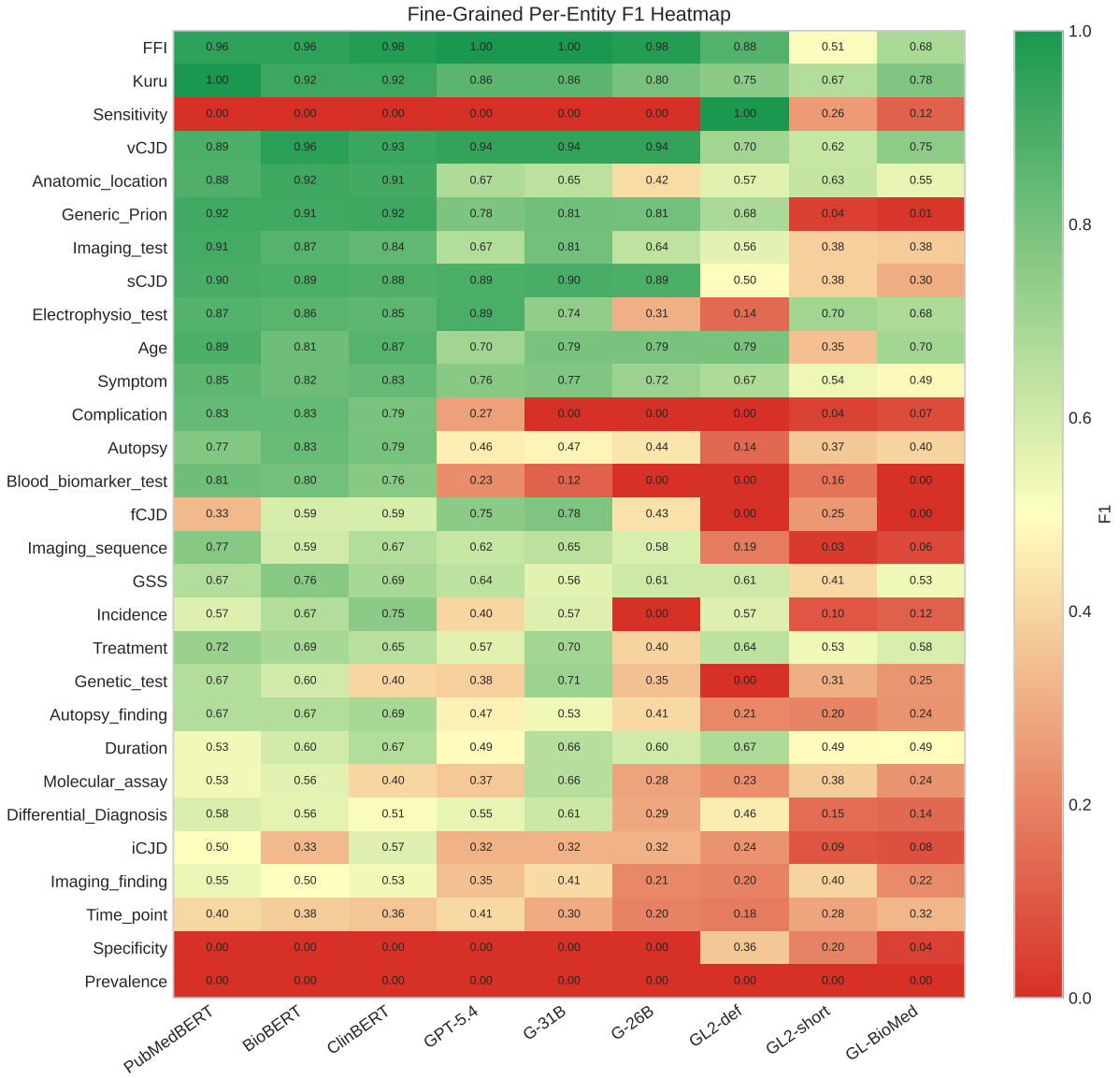


Figure 3: Per-type fine-grained F1 heatmap for the main supervised and zero-shot models on the test set.

type region also remains imperfectly separated, especially among *fCJD*, *iCJD*, *GSS*, *sFI*, and nearby subtype labels. These are largely fine-grained semantic confusions between clinically adjacent categories rather than arbitrary label noise. One recurring shared error is the prediction of *Generic_Prion* for *fCJD* and *iCJD*. For *fCJD*, a plausible explanation is that abstracts often realize the subtype as phrases such as "familial CJD" or "familial form of CJD," where the diagnostically important modifier appears before the base disease name. In these cases, the model appears to anchor on "CJD" while failing to consistently preserve the prefix that signals the familial subtype. The same pattern likely explains part of the *iCJD* → *Generic_Prion* confusion: the model often captures "CJD" but misses the preceding cue that indicates an iatrogenic or transmitted form. This error is amplified by longer compositional mentions such as "corneal transplant-related CJD," where the treatment-related phrase can be separated from the disease name; in such cases, the model may label "corneal transplant" as *Treatment* and the remaining "CJD" span as *Generic_Prion* rather than assigning the full mention to *iCJD*. Another plausible shared confusion is between *Complication* and *Symptom*. Some complication mentions are lexically symptom-like in isolation, but are labeled as *Complication* because they denote downstream consequences of disease progression rather than primary manifestations. For example, *dysphagia* can be a symptom in many settings, but in prion

Entity Type	Supervised			Zero-shot					
	PubMedBERT	BioBERT	ClinBERT	GPT-5.4	G-31B	G-26B	GL2-def	GL2-short	GL-BioMed
FFI	95.65	95.65	97.78	**100.00**	**100.00**	97.78	87.80	51.43	68.42
Kuru	**100.00**	92.31	92.31	85.71	85.71	80.00	75.00	66.67	77.78
Sensitivity	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	25.81	12.24
vCJD	88.70	**96.23**	92.73	94.34	94.34	94.34	70.37	62.26	74.77
Anatomic_location	88.28	**91.88**	91.18	67.19	64.79	41.61	56.81	62.83	55.35
Generic_Prion	91.77	91.45	**91.81**	78.38	80.95	80.69	68.33	3.86	1.35
Imaging_test	**90.79**	86.96	84.34	66.67	80.50	63.83	55.93	38.37	37.70
sCJD	**89.74**	89.03	87.90	88.89	89.66	88.89	49.54	38.37	30.19
Electrophysio_test	87.18	86.08	85.00	**89.19**	73.85	31.11	14.29	70.42	67.74
Age	**88.89**	81.48	86.79	70.18	79.25	79.25	79.25	34.92	69.84
Symptom	**84.95**	81.97	83.21	76.16	77.41	71.86	67.28	53.69	48.72
Complication	**83.33**	**83.33**	78.95	27.27	0.00	0.00	0.00	4.40	6.90
Autopsy	77.31	**82.76**	79.31	45.83	47.42	43.90	13.89	36.54	39.66
Blood_biomarker_test	**81.48**	80.49	75.86	22.64	12.20	0.00	0.00	16.33	0.00
fCJD	33.33	58.82	58.82	75.00	**77.78**	42.86	0.00	25.00	0.00
Imaging_sequence	**76.51**	59.42	66.67	62.30	64.62	58.33	18.60	2.50	6.20
GSS	66.67	**75.86**	68.97	64.29	56.25	60.61	60.87	40.91	53.33
Incidence	57.14	66.67	**75.00**	40.00	57.14	0.00	57.14	9.52	11.76
Treatment	**72.39**	69.01	65.43	56.94	69.57	39.64	64.43	53.27	58.45
Genetic_test	66.67	60.00	40.00	38.46	**71.43**	34.78	0.00	30.77	24.56
Autopsy_finding	66.67	66.67	**68.82**	46.67	53.12	40.68	21.28	20.34	24.18
Duration	53.06	59.57	66.67	49.02	65.98	59.77	**67.35**	49.02	49.48
Molecular_assay	52.78	56.00	40.00	36.92	**65.79**	28.07	23.08	38.20	23.91
Differential_Diagnosis	57.97	55.56	51.43	55.17	**61.29**	28.57	45.61	14.81	13.79
iCJD	50.00	33.33	**57.14**	31.58	31.58	31.58	24.00	9.23	7.69
Imaging_finding	**54.55**	50.38	53.03	34.92	41.18	21.43	20.25	39.71	21.90
Time_point	40.43	38.10	36.00	**40.78**	29.85	20.00	18.46	28.32	32.43
Specificity	0.00	0.00	0.00	0.00	0.00	0.00	**36.36**	20.00	3.92
Prevalence	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 16: Per-type fine-grained F1 scores for selected supervised and zero-shot models on the test set, sorted by the best score achieved for each entity type. Header abbreviations: ClinBERT = ClinicalBERT; GL2-def = GLiNER2-def; GL2-short = GLiNER2-short; GL-BioMed = GLiNER-BioMed; G-26B and G-31B = Gemma 4 26B and 31B.

disease abstracts it may appear as a late-stage secondary consequence of earlier neurological decline, which makes the boundary between the two labels difficult to recover from local context alone. Similarly, mentions such as "died" or "death" may be pulled toward *Symptom* or other outcome-like categories even though they are not direct symptomatic expressions of prion disease itself. Instead, they typically reflect the final downstream result of severe functional deterioration and accumulated health complications over the course of the disease.

For PubMedBERT, the strongest diagonal cells are concentrated on common, high-volume labels such as *Symptom*, *Anatomic_location*, *Imaging_test*, *Generic_Prion*, and *sCJD*. Its remaining errors mostly look like local mix-ups within semantically related label families rather than broad collapse across the label space. Gemma-4-31B, by contrast, is weaker on context-heavy diagnostic labels and on imaging-related distinctions; its diagonal is often present but noticeably lighter, suggesting less stable fine-grained label assignment even when an entity span is recovered.

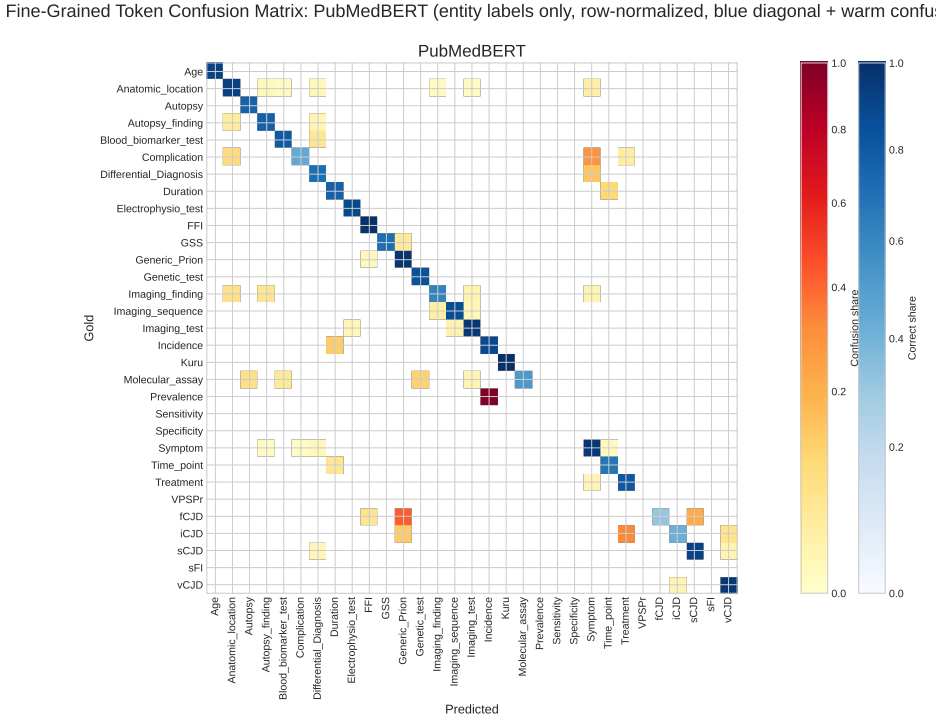
This entity-only view should also be read together with the full confusion analysis that includes *O*. Because these matrices exclude non-entity predictions, they likely understate Gemma-4-31B's main weakness, namely omission of the entity altogether. In other words, PubMedBERT appears stronger and more balanced, with errors concentrated within semantically adjacent labels, whereas Gemma-4-31B shows reasonable zero-shot behavior but is less reliable on subtle distinctions and is likely more omission-heavy in the full setting.

At the label level, PubMedBERT remains strong on common clinical categories, including *Generic_Prion* (0.918), *Imaging_test* (0.908), *sCJD* (0.897), *vCJD* (0.887), *Anatomic_location* (0.883), and *Symptom* (0.850). Gemma-4-31B is still strong on a narrower set of distinctive labels,

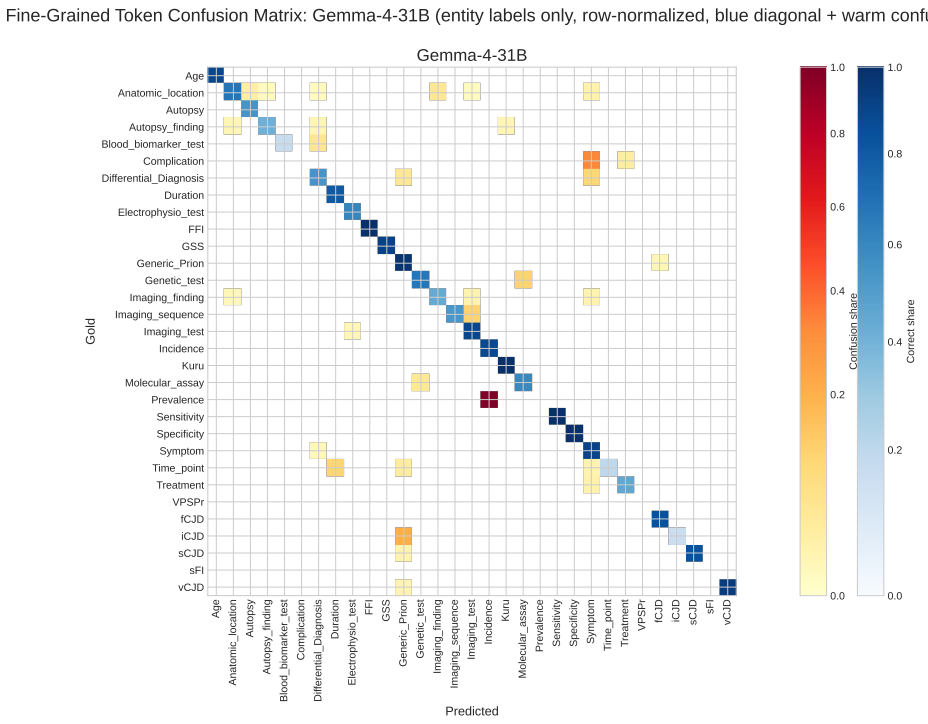
including *Generic_Prion* (0.917), *FFI* (0.977), *Kuru* (0.923), *Age* (0.792), *Imaging_test* (0.781), and *Symptom* (0.773), but its performance drops sharply on rarer or more context-dependent categories. Its weakest flat F1 values include *Complication* (0.000), *Prevalence* (0.000), *Sensitivity* (0.000), *Specificity* (0.000), *Blood_biomarker_test* (0.172), *iCJD* (0.286), *Time_point* (0.317), and *Imaging_finding* (0.321). PubMedBERT also shows weak spots, but they are milder and are concentrated in rare epidemiology labels and subtype distinctions: *Prevalence* (0.000), *Sensitivity* (0.000), *Specificity* (0.000), *fCJD* (0.333), *Time_point* (0.404), *iCJD* (0.500), and *Molecular_assay* (0.528). Overall, these figures reinforce the main result: PubMedBERT is the better model for PrionNER because it maintains Gemma-4-31B-level precision while recovering many more entities, whereas Gemma-4-31B remains the strongest zero-shot model but is still limited primarily by recall.

D.4 Per-label Annotation Agreement

Table 17 provides the full per-label agreement breakdown for the test split. Agreement is highest for common and semantically distinctive disease labels such as *vCJD*, *GSS*, *FFI*, and *sCJD*, while lower agreement is concentrated in sparse or boundary-sensitive categories such as *Complication*, *Prevalence*, *Incidence*, and *Imaging_finding*.



(a) PubMedBERT



(b) Gemma-4-31B

Figure 4: Entity-only fine-grained confusion matrices for the strongest supervised model (PubMedBERT) and the strongest zero-shot model (Gemma-4-31B).

Label	F1	Matches	Ann. 1	Ann. 2	Union	Jaccard
Complication	0.00	0	12	2	14	0.00
Prevalence	0.00	0	2	4	6	0.00
Incidence	40.00	1	2	3	4	25.00
Imaging_finding	47.19	21	49	40	68	30.88
Autopsy_finding	51.61	16	37	25	46	34.78
Blood_biomarker_test	52.63	20	32	44	56	35.71
Time_point	60.61	20	34	32	46	43.48
Treatment	63.01	46	58	88	100	46.00
Differential_Diagnosis	63.64	28	36	52	60	46.67
Autopsy	76.54	31	43	38	50	62.00
Age	76.92	20	26	26	32	62.50
Duration	77.65	33	40	45	52	63.46
iCJD	78.26	9	11	12	14	64.29
Imaging_sequence	80.00	50	66	59	75	66.67
Genetic_test	80.00	4	4	6	6	66.67
Symptom	81.00	194	235	244	285	68.07
Specificity	83.33	5	5	7	7	71.43
Electrophysio_test	84.51	30	34	37	41	73.17
Sensitivity	85.71	6	6	8	8	75.00
Molecular_assay	87.64	39	47	42	50	78.00
Anatomic_location	90.30	149	159	171	181	82.32
Imaging_test	92.31	60	64	66	70	85.71
Kuru	93.33	7	7	8	8	87.50
Generic_Prion	94.83	220	234	230	244	90.16
fCJD	95.24	10	11	10	11	90.91
sCJD	95.54	75	79	78	82	91.46
FFI	95.65	22	22	24	24	91.67
vCJD	96.55	56	58	58	60	93.33
GSS	96.77	15	16	15	16	93.75

Table 17: Per-label annotation agreement under exact label-and-span matching on the 70-abstract test split, with one abstract excluded from the final agreement comparison. Ann. 1 and Ann. 2 denote the number of entities annotated by Annotator 1 and Annotator 2, respectively, Matches denotes exact label-and-span matches, and Union is the size of the union of annotated entities for that label.

Coarse-grained Type	Fine-grained Type	Train Count	Train %	Test Count	Test %
Generic_Prion	Generic_Prion	741	15.92	233	14.12
Sporadic_Prion	sCJD	86	1.85	80	4.85
Sporadic_Prion	sFI	3	0.06	0	0.00
Sporadic_Prion	VPSPr	0	0.00	0	0.00
Familial_Prion	fCJD	26	0.56	10	0.61
Familial_Prion	GSS	32	0.69	16	0.97
Familial_Prion	FFI	39	0.84	22	1.33
Acquired_Prion	vCJD	102	2.19	55	3.33
Acquired_Prion	iCJD	32	0.69	16	0.97
Acquired_Prion	Kuru	28	0.60	7	0.42
Differential_Diagnosis	Differential_Diagnosis	188	4.04	36	2.18
Symptom	Symptom	791	16.99	248	15.03
Test_name	Imaging_test	261	5.61	75	4.55
Test_name	Electrophysio_test	173	3.72	38	2.30
Test_name	Blood_biomarker_test	64	1.37	39	2.36
Test_name	Genetic_test	16	0.34	8	0.48
Test_name	Molecular_assay	23	0.49	45	2.73
Test_name	Autopsy	199	4.27	64	3.88
Sequences	Imaging_sequence	117	2.51	77	4.67
Findings	Imaging_finding	153	3.29	71	4.30
Findings	Autopsy_finding	215	4.62	42	2.55
Anatomic_location	Anatomic_location	646	13.88	223	13.52
Age	Age	137	2.94	29	1.76
Treatment	Treatment	126	2.71	88	5.33
Complication	Complication	78	1.68	19	1.15
Time	Duration	217	4.66	49	2.97
Time	Time_point	152	3.27	47	2.85
Stats	Sensitivity	0	0.00	4	0.24
Stats	Specificity	0	0.00	4	0.24
Stats	Prevalence	4	0.09	1	0.06
Stats	Incidence	6	0.13	4	0.24
Total		**4655**	**100.00**	**1650**	**100.00**

Table 18: Full fine-grained entity distribution in the PrionNER train and test splits. Percentages are computed separately within the train and test sets over schema-defined fine-grained entity mentions only.

E Additional Reference Tables

E.1 Full Fine-grained Entity Distribution

E.2 Top Surface Forms by Entity Type

Table 19 reports the 10 most frequent normalized surface forms for each entity type. Numbers in parentheses after each surface form indicate mention counts, and the left-column notation "U./M." denotes total unique surface forms and total mentions for that entity type.

Table 19: Top 10 normalized surface forms for each entity type in the combined dataset.

Entity Type (U./M.)	Top Surface Forms (mentions)
Symptom (524 / 1279)	dementia (79); myoclonus (79); ataxia (47); rapidly progressive dementia (36); psychiatric symptoms (24); akinetic mutism (16); progressive dementia (16); myoclonic jerks (15); aggression (15); neurological symptoms (14)
Anatomic_location	

Continued on next page

Entity Type (U./M.)	Top Surface Forms (mentions)
(276 / 837)	brain (48); basal ganglia (41); white matter (29); cortical (26); cerebellum (23); cerebellar (23); thalamus (22); striatum (21); cerebral cortex (19); pulvinar (18)
Duration	
(201 / 290)	7 months (8); within a year (7); 12 months (7); four months (6); 13 months (6); within one year (5); 4 months (5); 3 months (5); 5 months (5); disease duration (4)
Time_point	
(181 / 248)	age at onset (5); at onset (5); 1996 (5); after onset (5); in 1996 (5); 1985 (4); at the onset (4); the end of the third stage (4); clinical onset (3); early in the course (3)
Imaging_finding	
(161 / 237)	pulvinar sign (17); brain atrophy (13); hyperintensity (11); high signal intensity (7); atrophy (6); hyperintense signal abnormalities (4); signal intensity abnormalities (4); high signal (4); hyperintensities (3); cortical atrophy (3)
Autopsy_finding	
(155 / 276)	neuronal loss (22); spongiform change (20); spongiosis (9); spongiform changes (8); gliosis (8); status spongiosus (8); astrocytosis (7); neuronal degeneration (5); spongiform encephalopathy (5); kuru plaques (5)
Differential_Diagnosis	
(136 / 278)	alzheimer's disease (17); dlb (16); dementia (14); ad (12); stroke (9); neurodegenerative disorder (7); non-prion disorders (6); insomnia (5); dementias (4); thalamic dementia (4)
Treatment	
(131 / 241)	palliative care (23); quinacrine (16); antipsychotics (9); quetiapine (6); haloperidol (6); treatment (4); anesthesia (4); supportive care (4); hospice care (4); dura mater graft (4)
Age	
(124 / 182)	61-year-old (5); elderly (5); 59-year-old (5); 49-year-old (5); 54-year-old (4); 65-year old (3); 48-year-old (3); 70-year-old (3); 75-year-old (3); 58-year-old (3)
Imaging_test	
(100 / 383)	mri (82); magnetic resonance imaging (37); ct (28); spect (27); mr imaging (17); mr (13); mr images (13); ct scan (9); positron emission tomography (7); brain mri (7)
Generic_Prion	
(85 / 1219)	cjd (576); creutzfeldt-jakob disease (327); prion disease (50); bse (33); prion diseases (30); bovine spongiform encephalopathy (24); spongiform encephalopathy (16); creutzfeldt-jakob disease (14); creutzfeldt-jacob disease (9); subacute spongiform encephalopathy (6)
Autopsy	
(74 / 271)	autopsy (62); brain biopsy (36); necropsy (20); biopsy (13); neuropathological (11); neuropathological examination (9); neuropathologically (7); pathological (5); pathologically (5); histologically (5)
Imaging_sequence	
(63 / 241)	dwi (34); flair (34); diffusion-weighted (15); diffusion-weighted imaging (12); t2-weighted (12); t2 (11); diffusion-weighted images (9); dw (8); adcs (8); dw images (8)
Electrophysio_test	
(49 / 247)	eeg (119); electroencephalogram (29); electroencephalographic (15); electroencephalography (10); eegs (8); electroencephalograms (6); erg (4); polysomnography (4); br (3); electroencephalographic findings (3)
Blood_biomarker_test	
(44 / 140)	csf (57); cerebrospinal fluid (14); blood tests (9); csf analysis (6); csf tau protein (4); csf virology (3); biomarkers (3); csf tau-pt181 (2); csf studies (2); serum workup (2)
Molecular_assay	
(40 / 111)	cdi (16); csf rt-quic (6); immunohistochemistry (6); ihc (6); western blot (5); immunocytochemistry (5); molecular analysis (4); densitometric analysis (4); protein assay (3); csf rt-qulc (3)
sCJD	
(26 / 241)	scjd (90); sporadic cjd (52); sporadic creutzfeldt-jakob disease (29); sporadic (17); vv2 (8); heidenhain variant (4); creutzfeldt-jakob disease (4); sporadic form (4); mv1 (4); mm1 (4)
Complication	
(22 / 52)	death (20); died (7); myocarditis (3); deaths (2); bronchopneumonia (2); fatal (2); dysphasia (1); loss of independence (1); dead (1); acute myocarditis (1)
iCJD	
(19 / 51)	iatrogenic cjd (8); iatrogenic creutzfeldt-jakob disease (8); dcjd (8); icjd (5); iatrogenic (4); iatrogenic forms (3); iatrogenic cases (2); dural graft associated cjd (2); iatrogenic transmission (1); iatrogenic transmission of cjd (1)
GSS	

Continued on next page

Entity Type (U./M.)	Top Surface Forms (mentions)
(18 / 63)	gss (19); gerstmann-sträussler-scheinker disease (14); gss102 (6); gerstmann-straussler-scheinker disease (4); gss105 (4); gssd (3); gerstmann-straussler-scheinker syndrome (2); gerstmann-sträussler-scheinker's disease (1); gerstmann-sträussler-scheinker's syndrome (1); gerstmann-strässler-scheinker's syndrome (1)
vCJD	
(18 / 218)	vcjd (104); nvcjd (37); variant creutzfeldt-jakob disease (26); variant cjd (26); nv-cjd (4); acquired (4); variant creutzfeldt-jakob (2); infectious (2); new variant cjd (2); variant (2)
fCJD	
(16 / 47)	familial (9); familial cjd (8); inherited prion disease (6); inherited (4); fcjd (3); genetic cjd (3); genetic (3); hereditary cjd (2); cjd178 (2); genetic forms (1)
Genetic_test	
(15 / 26)	molecular genetic analysis (5); genetic testing (3); genetic analysis (2); prp gene analysis (2); restriction-enzyme analysis (2); analysing dna (2); prion gene analysis (2); genetic examination (1); genotyping (1); genetic tests (1)
Incidence	
(11 / 12)	high incidence (2); 0.06% in 2023 (1); 0.10% in 2024 (1); 1 to 2 cases per million people per year (1); annual incidence of 0.37 cases/million (1); incidence (1); annual incidence of 0.5-1.5 cases of cjd per million (1); incidence of approximately .5-1 new cases per million population per year (1); incidence 1 in 1 000 000 (1); 1 to 2 cases per million people per year. (1)
Prevalence	
(9 / 10)	1–2 people per million annually (2); 1–2/million/year (1); one case per million people per year (1); one in one million (1); 1 to 2 cases per million people per year (1); 35% (1); 1–2 people per million annually. (1); 0.06% (1); 0.10% (1)
FFI	
(7 / 85)	ffi (44); fatal familial insomnia (33); ffi-1 (4); ffi-2 (1); fatal familiar insomnia (1); met-met subtype (1); fatal insomnia (1)
Sensitivity	
(7 / 15)	sensitivity (4); sensitivity of 100% (2); sensitivity of 87% (2); sensitivity (96%) (2); 91% sensitive (2); sensitivity higher (2); sensitive (1)
Specificity	
(5 / 12)	specificity (4); specificity of 92% (2); specificity of 97% (2); specificity (97%) (2); 95% specific (2)
Kuru	
(3 / 43)	kuru (40); kuru plaques (2); kuru type (1)
sFI	
(3 / 3)	ffi-1 (1); ffi-2 (1); fatal familial insomnia (1)

Evaluation of Multilingual Text Simplification for the Mental Health Domain: Exploring Small Language Models

Olga Pelloni[1], Sandra Anna Just[2,3], Lars Ailo Bongo[1]

[1]Department of Computer Science, UiT The Arctic University of Norway,
[2]Department of Clinical Medicine, UiT The Arctic University of Norway,
[3]Department of Psychiatry and Neurosciences, Charité—Universitätsmedizin, Berlin, Germany
{olga.pelloni, sandra.a.just, lars.ailo.bongo}@uit.no

Abstract

Individuals with particular mental health disorders may find it difficult to learn about their own condition. Therefore, efforts have been made to create materials that explain complex medical information in simpler words, which are also beneficial for caregivers and others. However, text simplification is commonly done in English and only sporadically in other languages. In this study, we explore potential ways for language-agnostic medical text simplification for the mental health domain. Our approach is to simplify the ICD-11 articles on primary psychotic disorders in English, German and French, using small LMs and various metrics for evaluating different aspects of the texts: lexical complexity and readability. Our results show that acceptable texts were produced only in English, and that a joint analysis of Measure of Textual Lexical Diversity (MTLD) and Flesch Reading Ease (FRE) provides the most insight, capturing both the best outcomes and signaling different types of issue. The study is preliminary and requires further investigation.

1 Introduction

As with any professional or technical domain, medical texts tend to be difficult to understand for non-specialists. Clinical NLP explores methods for different transformations of medical texts (Hasan and Farri, 2019), including *paraphrase generation* and *text simplification*, where the latter is a more complex task even in English (Basu et al., 2023).

Simplification of medical texts in languages other than English remains understudied. Some research on French and Romanian focuses primarily on paraphrase generation and shows good results when using language- and domain-specific models (Buhnila and Todirascu, 2024; Buhnila et al., 2024; Buhnila, 2025), whereas research on text simplification in Polish and some Indian languages, Arabic and Swahili reports that general purpose LLMs, such as GPT-4, provide better results than smaller fine-tuned models (Grzybowski et al., 2025; Gangavarapu, 2024).

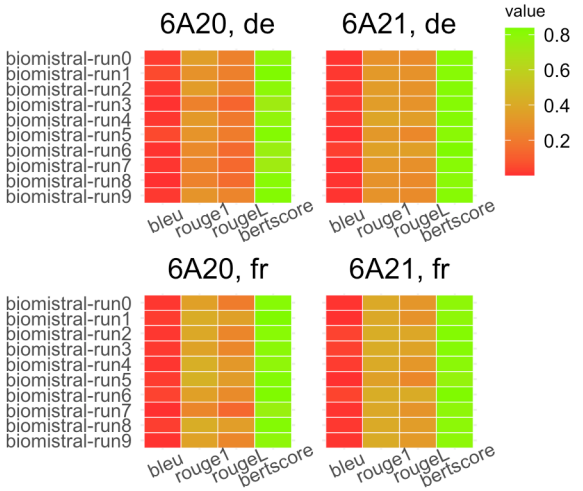


Figure 1: Typically used similarity metrics BLEU, ROUGE and BERTScore do not reflect fine-grained differences in simplified texts when comparing different languages, model runs and ICD-11 articles, although the resulting texts vary greatly in terms of fluency and simplicity (our results).

In contrast to text simplification in primary medicine, the mental health domain requires further attention to potential cognitive impairments in some mental health disorders, neutral and non-stigmatizing tone, simplifying both words and sentences without making text unnatural. These nuances can be challenging to account for in multilingual settings, even when translating medical texts manually (Natoli, 2026).

Given the ambitious nature of the outlined task, we conduct a preliminary study to investigate evaluation methods capable of detecting various issues in a multilingual setting. In this study, we explore small models, intentionally including some negative results (poor-quality texts), to focus on identifying key metrics that highlight text properties of the best and worst simplified texts across 3 languages.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 464–474
July 3-4, 2026 ©2026 Association for Computational Linguistics

Small models are chosen for greater reproducibility in a metric exploration study.

2 Related Work

The need for reading materials in simplified language has long been recognized, especially in healthcare contexts. Different lay persons are considered beneficiaries of so-called Easy Language, and various practices for it have been established across countries for manual adaptation of texts (Nomura et al., 2010; Lindholm and Vanhatalo, 2021).

Despite these efforts, automatic methods for clinical text simplification remain heavily focused on English, with only limited work on other European languages (Buhnila and Todirascu, 2024; Buhnila et al., 2024; Buhnila, 2025; Grzybowski et al., 2025; Scholz and Wenzel, 2025) and even less beyond Europe (Gangavarapu, 2024).

Automatic simplification of medical texts raises ethical considerations, such as the risk of meaning distortion. (Gooding, 2022) This is even more important when targeting texts in the psychiatric domain (Aich et al., 2025) and audiences with mental health disorders (Gralton et al., 2010), given the long history of stigmatization (Rössler, 2016). Evaluation of the resulting texts typically relies on metrics coming from machine translation (MT) that measure similarity between the original and the simplified texts on the level of word tokens (Papineni et al., 2002; Lin, 2004) and on semantic level (Zhang et al., 2019), which provide only limited insight on the quality of the simplified texts (Figure 1). The Flesch Reading Ease metric (Flesch, 1948) is widely used to assess readability, while the other readability metrics and other properties of text, such as lexical complexity, are typically overlooked.

Our preliminary study addresses these gaps by analyzing 3 lexical complexity and 9 readability metrics to identify more informative and potentially language-agnostic evaluation methods.

3 Data & Methods

We extract five articles from ICD-11 (Harrison et al., 2021) using ICD API[1], containing definitions of primary psychotic disorders: 6A20–6A24 including all the sub-definitions (6A20.0, 6A20.1, etc.). The concatenated definitions serve as input to the following small language models, fine-tuned for text simplification and/or for multilingual use:

- **LingConv** (Elgaar and Amiri, 2025): a text conversion tool[2] based on a T5-Base (Raffel et al., 2020) model intended for general paraphrase generation in English;
- **PRISM** (Thompson and Post, 2020): a transformer-based MT model adapted for general paraphrase generation, trained on 39 languages;
- **Ascle** (Yang et al., 2024): a toolkit for medical text generation, incl. a text simplification function. The toolkit includes 3 models (BigBird-Pegasus Large (Zhang et al., 2020); BART Large (Lewis et al., 2020) and BioBART v2 Base (Yuan et al., 2022)), each in 2 versions fine-tuned on different medical data, primarily in English.
- **BioMistral 7B** (Labrak et al., 2024): a Mistral 7B Instruct model (Jiang et al., 2023) fine-tuned for biomedical domain in 9 languages. We use the BioMistral SLERP model[3] with 4-bit GPTQ quantization (Frantar et al., 2022).

All models except BioMistral were given the ICD-11 articles as input, while for BioMistral we provided a simple prompt: "Rewrite this medical definition for a layman person. Keep it as close to the original meaning as possible, just use simpler words and syntactic constructions."; we manually translated it into German and French.

We test all the models on English texts first and calculate the following evaluation metrics:

- **Similarity**: BLEU, ROUGE 1 and ROUGE L, BERTScore. These are standard metrics in MT for assessing the preservation of the n-gram combinations and the underlying meaning;
- **Lexical complexity**: TTR (type-token ratio), MATTR (moving-average type-token ratio) (Covington and McFall, 2010), MTLD (measure of textual lexical diversity) (McCarthy and Jarvis, 2010). All of these measures are based on the ratio of unique words (types) and all word instances (tokens) in a given text;
- **Readability**: Flesch Reading Ease (FRE), Flesch-Kincaid grade level (FKGL), Automated Readability Index (ARI) (Kincaid et al.,

[1] https://icd.who.int/icdapi

[2] https://mohdelgaar-lingconv.hf.space/

[3] https://huggingface.co/LoneStriker/BioMistral-7B-SLERP-GPTQ

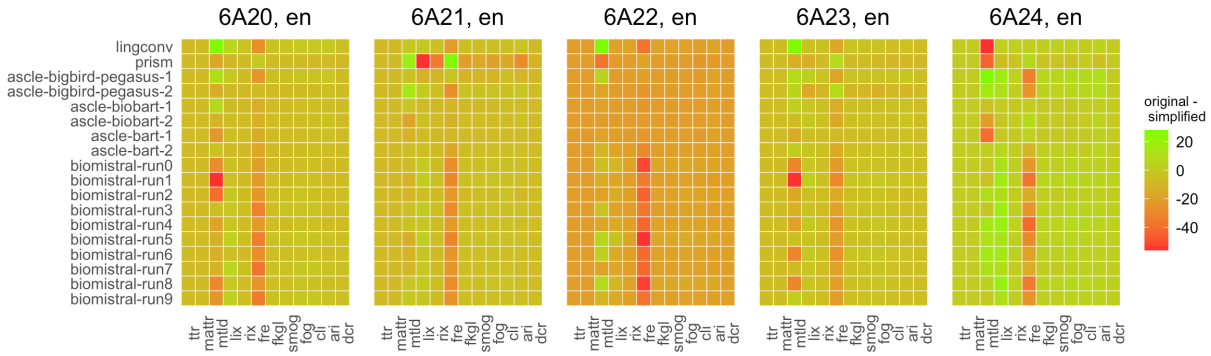


Figure 2: Difference in complexity and readability metrics across 9 models and 5 ICD-11 articles in English.

1975), SMOG index (SMOG) (Mc Laughlin, 1969), Gunning Fog index (Fog) (Gunning, 1952), Coleman–Liau Index (CLI) (Coleman and Liau, 1975), Dale-Chall Readability (DCR) (Chall and Dale, 1995), Läsbarhetsindex (LIX) (Björnsson, 1968) and Readability Index (RIX) (Anderson, 1983). These metrics capture length of words, sentences; some of them also check a set of familiar words (e.g. Dale-Chall).

Given the results on English, we manually evaluate the simplified texts from each model and identify the most aligned metrics. The reduced set of metrics is further used to evaluate the results on German and French, taking into account that some of the readability metrics cannot be easily transferred to languages other than English.[4] In addition, three human annotators evaluated BioMistral results in English.

We use textstat (Bansal and Aggarwal, 2026) and LexicalRichness (Shen, 2022) Python libraries to calculate all the readability and lexical complexity metrics. In particular, we observe the *differences* in these metrics between the original and simplified texts, so that rather than relying on absolute metric values, we assess the *reduction in complexity*, which allows us to do better comparisons cross-linguistically. Our code is open sourced and available at: `https://github.com/uit-hdl/multilingual-text-simplification-icd-11`

4 Results & Discussion

Figure 2 shows the differences in complexity and readability metrics measured on English texts. The general paraphrasing models LingConv and PRISM produce the worst texts, but in different ways: LingConv tends to produce just one short and simple sentence (e.g. "Schizophrenia is known as a disease."), while PRISM tends to keep sentences very close to the original definitions ("Schizophrenia can be characterized by disorders").

Rank	#1	#2	#3
Best (1)	4, 0	**8**, 7	0
Good (2)	**9**, **8**	**9**, 6	1, 2
Medium (3)	7, 6	2, 3	3, 7, 8, 9
Poor (4)	**5**, **1**	4, **5**	4, 6
Worst (5)	3, 2	**1**, 0	**5**

Table 1: Human experts' rankings of the texts in English, BioMistral model's runs 0–9, grouped into categories from 1 to 5. We mark runs with $\geq$2 annotator agreement and no opposing third (Cohen's $\kappa \approx 0$).

Thus, the first two rows stand for extreme cases, where LingConv is too simple and PRISM is too complex. Ascle models produced texts of mediocre quality, somewhat close to PRISM but not as complex (e.g. "Schizophrenia is a mental illness that is characterized by changes").[5] BioMistral model produced the most natural simplified texts when given 6A20 article as input compared to the other models (e.g. "Schizophrenia is a mental illness that affects how you think, feel, and behave").

Observing the differences in metrics, we notice that MTLD (3rd column in Figure 2) and FRE (6th column) are the most prominent features[6]. The best simplified texts of the 6A20 article in English produced by BioMistral are characterized by increase in MTLD (in red) with a *simultaneous* increase in FRE (in red). The only results with substantial increase of these two metrics were produced by BioMistral when processing 6A20 and 6A23

[4]The library textstat provides adaptations of FRE and Fog to several languages: https://pypi.org/project/textstat/.

[5]See original texts in Appendix, Table 2.

[6]Note that some of the metrics cannot show big difference in values by nature, e.g. based on grade level; we observed each of these metrics separately and did not find any consistent patterns with respect to the quality of the models' output.

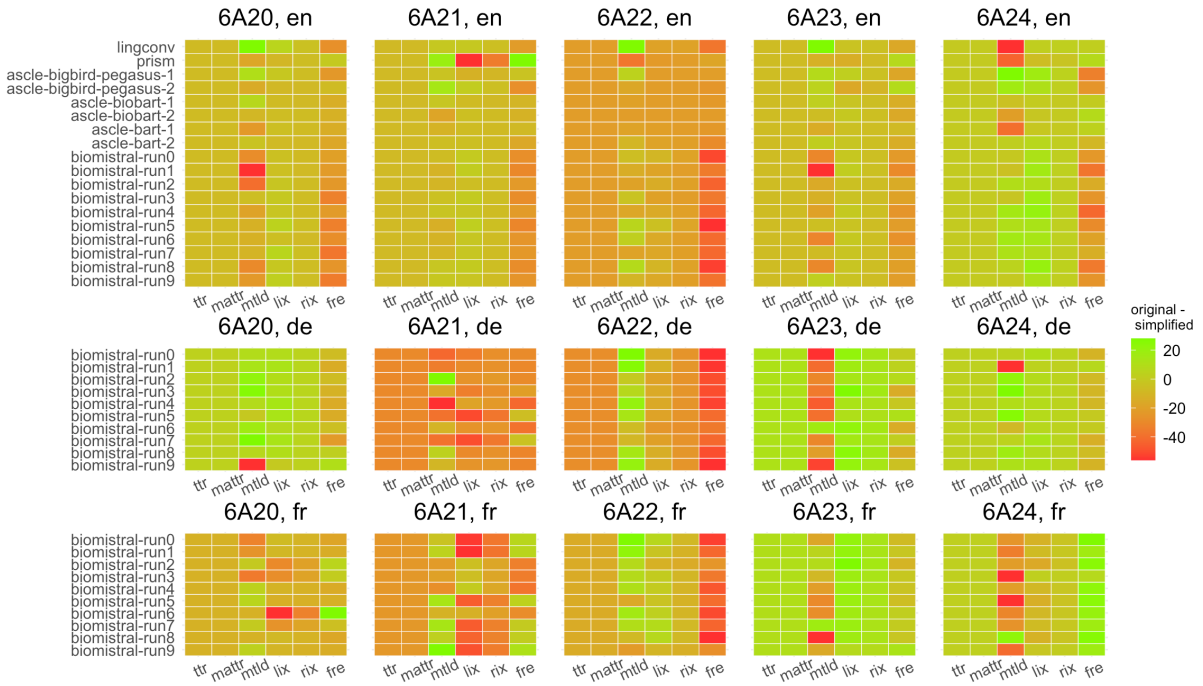


Figure 3: Difference in complexity and readability metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

articles.

The other models mostly produce texts which either show a decrease in MTLD (in green) and increase in FRE (in red) or vice versa, or no change compared to the original texts.

We did an additional check by three human experts to evaluate the texts produced by BioMistral in English for 6A20 definition to see if the fluctuations in MTLD and FRE align with human judgements (Table 1). While the human experts reported that generally the texts look similar to each other, they could identify some best and worst results within the BioMistral runs: at least two of the experts mark runs 8, 9 as the best and 1, 5 as the worst. Best results have a moderate increase in MTLD (in the range of 27–37), unlike the worst texts, where the increase in MTLD is very high (82 in run 1) or very low (8 in the run 5) (see examples in Appendix, Table 3).

Figure 3 shows the differences in complexity metrics across English, German and French. We only present results of the BioMistral model, as the other multilingual model PRISM did not produce any paraphrases neither in German nor in French; in fact, some English outputs contained mixed expressions with French or Spanish (e.g. "caractérisated"). BioMistral in German and French struggles more than in English: texts mostly remain complex, contain grammatical mistakes and are not smooth. This is reflected by increase in MTLD (in red) and decrease in FRE (in green) and vice versa (see examples in Appendix, Table 4, 5).

Some poor-quality results might be due to the inherent properties of the input text, e.g. 6A22 is a short article, thus, not much transformation is happening in all three languages: FRE increases as expected, but MTLD consistently decreases. This is also noticeable in differences of the number of word tokens and types (see Appendix, Figure 4).

5 Conclusion

In this study, we explored 3 lexical complexity and 9 readability metrics to evaluate multilingual text simplification in the mental health domain. We find that a combination of differences in MTLD and FRE helps evaluating simplified texts: increase of FRE means that words become shorter, and increase in MTLD makes the lexical diversity higher, while decrease in MTLD results in repetitive words. Increase in MTLD and decrease in FRE signify the opposite: words become longer and there is high lexical diversity, making texts too dense. In addition, small LMs worked well only on English, while German and French resulted in poor-quality texts. We suggest exploring these metrics for medical text simplification further with LLMs and using them as a pre-check for human evaluation.

Limitations

This is a preliminary study, and we acknowledge that stronger claims could be made if we had extended our analysis to more models, languages and text samples. We report mostly negative results in terms of the quality of the simplified texts; however, we intentionally limit this study to small models to allow for better control and exploration of the evaluation metrics. We recognize that higher-quality texts and more effective use of the metrics would likely be achieved when using large language models.

References

Ankit Aich, Tingting Liu, Salvatore Giorgi, Kelsey Isman, Ruhshana Bobojonova, Lyle Ungar, and Brenda Curtis. 2025. Language models in digital psychiatry: challenges with simplification of healthcare materials. *NPP—Digital Psychiatry and Neuroscience*, 3(1):10.

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Shivam Bansal and Chaitanya Aggarwal. 2026. Textstat: Python readability statistics library. https://github.com/textstat/textstat.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14093–14101.

Carl-Hugo Björnsson. 1968. *Läsbarhet: hur skall man som författare nå fram till läsarna?* Bokförlaget Liber.

Ioana Buhnila. 2025. Explain this medical term in my language: A case study of small language models for medical paraphrase generation. In *3rd UniDive Workshop*, pages 1–4, Budapest, Hungary.

Ioana Buhnila, Aman Sinha, and Mathieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.

Ioana Buhnila and Amalia Todirascu. 2024. LARGEMED: A resource for identifying and generating paraphrases for French medical terms. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 141–151, Torino, Italia. ELRA and ICCL.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Michael A Covington and Joe D McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100.

Mohamed Elgaar and Hadi Amiri. 2025. LingConv: An interactive toolkit for controlled paraphrase generation with linguistic attribute control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–51, Suzhou, China. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Agasthya Gangavarapu. 2024. Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions. *arXiv preprint arXiv:2404.08705*.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Ernest Gralton, Marilyn Sher, and Coro Drew Lopez. 2010. Information and readability issues for psychiatric patients: e-learning for users. *The Psychiatrist*, 34(9):376–380.

Łukasz Grzybowski, Jakub Pokrywka, Michał Ciesiółka, Jeremi Ignacy Kaczmarek, and Marek Kubis. 2025. Polish-English medical knowledge transfer: A new benchmark and results. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9042–9063, Suzhou, China. Association for Computational Linguistics.

R. Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

James E Harrison, Stefanie Weber, Robert Jakob, and Christopher G Chute. 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC medical informatics and decision making*, 21(Suppl 6):206.

Sadid A Hasan and Oladimeji Farri. 2019. Clinical natural language processing with deep learning. In *Data Science for Healthcare: Methodologies and Applications*, pages 147–171. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Technical report, Naval technical training command.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Camilla Lindholm and Ulla Vanhatalo. 2021. *Handbook of easy languages in Europe*. Frank & Timme.

G Harry Mc Laughlin. 1969. SMOG grading – a new readability formula. *Journal of reading*, 12(8):639–646.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Adam P Natoli. 2026. Leveraging artificial intelligence to linguistically compare test translations: A methodological introduction and demonstration. *Assessment*, 33(2):163–177.

Misako Nomura, Gyda Skat Nielsen, and Bror Tronbacke. 2010. Guidelines for easy-to-read materials. *IFLA professional reports*, (120):0_1.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Wulf Rössler. 2016. The stigma of mental disorders: A millennia-long history of social exclusion and prejudices. *The EMBO Reports*, 17(9):1250–1253.

Karen Scholz and Markus Wenzel. 2025. Evaluating readability metrics for German medical text simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness. `https://github.com/LSYS/lexicalrichness`. DOI: 10.5281/zenodo.6607007.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*, Online. Association for Computational Linguistics.

Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, and 1 others. 2024. Ascle—a python natural language processing toolkit for medical text generation: development and evaluation study. *Journal of medical Internet research*, 26:e60601.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

A Appendix

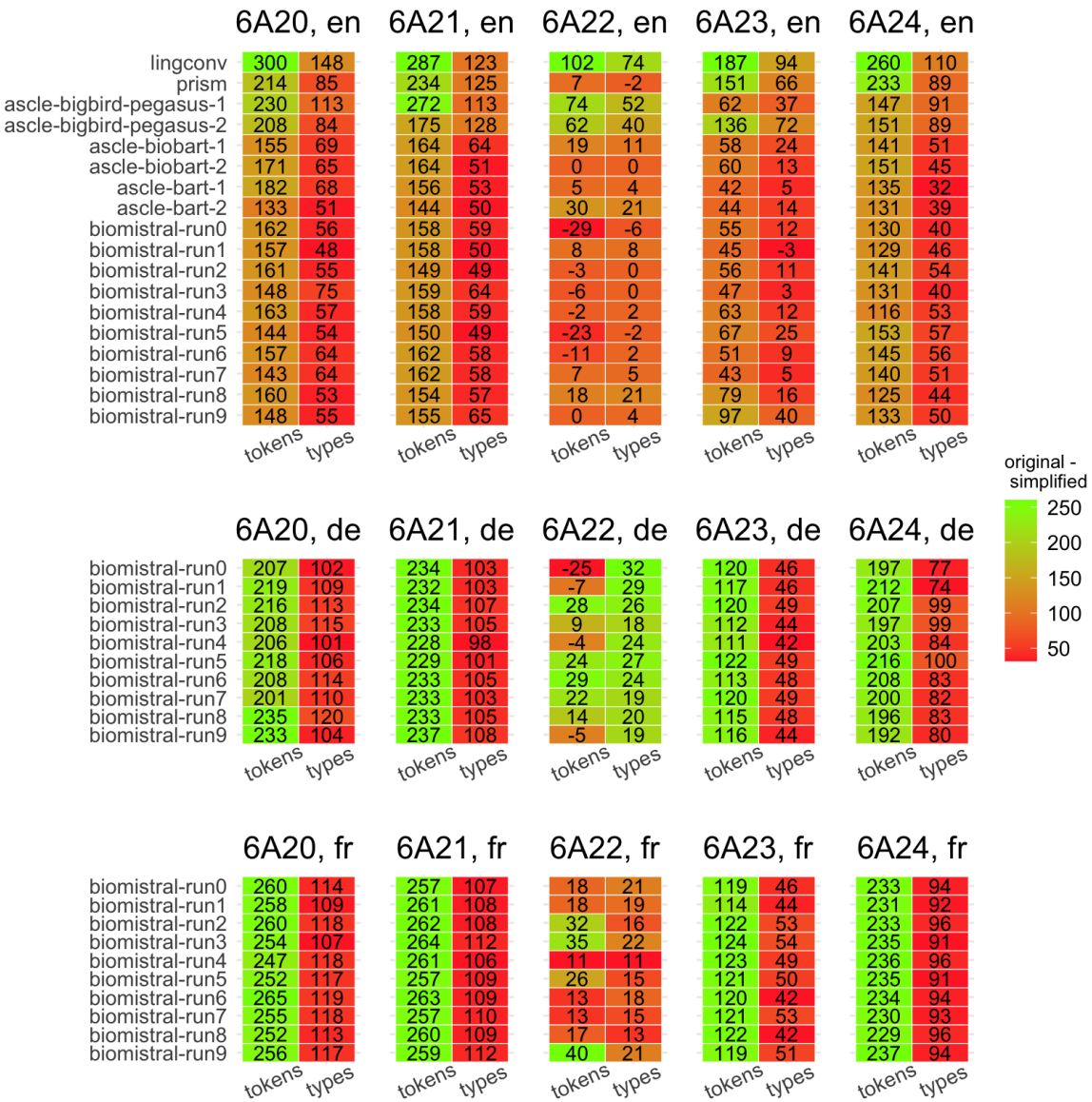


Figure 4: Difference in number of word tokens and word types (unique words) between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

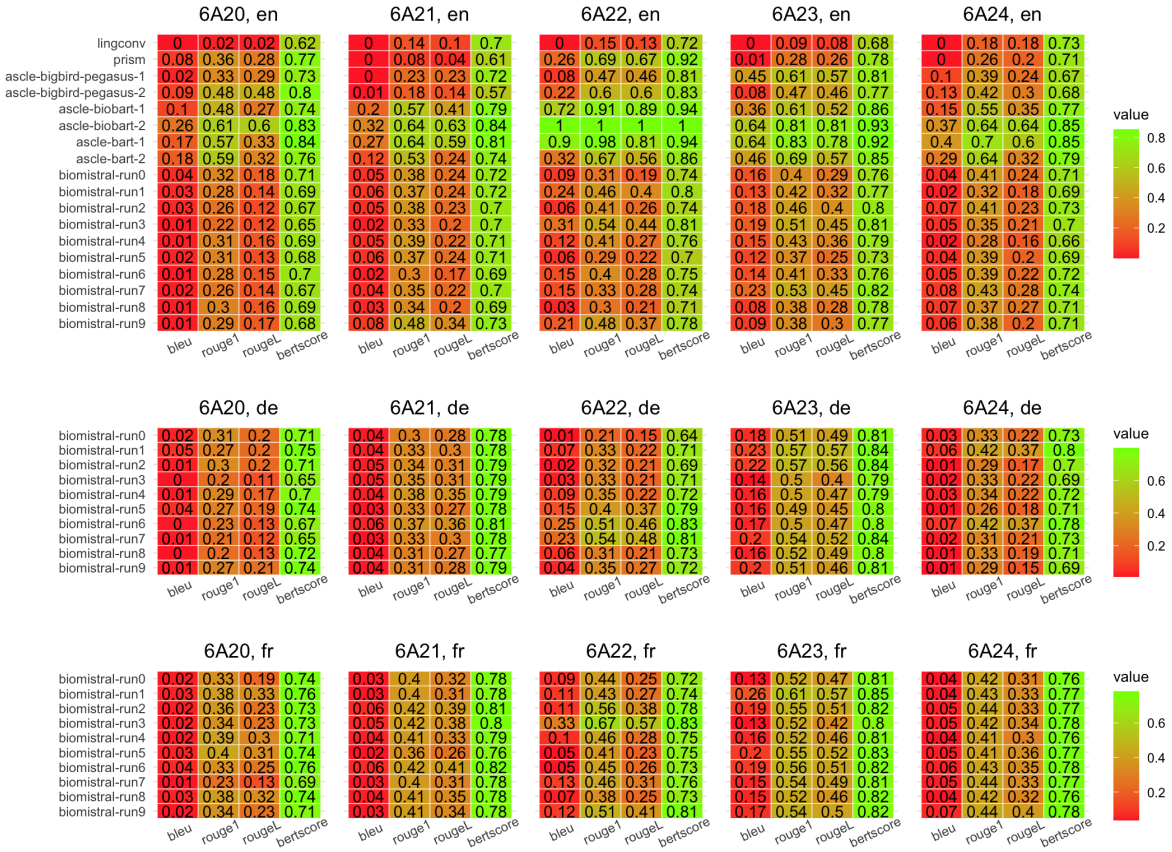


Figure 5: Similarity metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

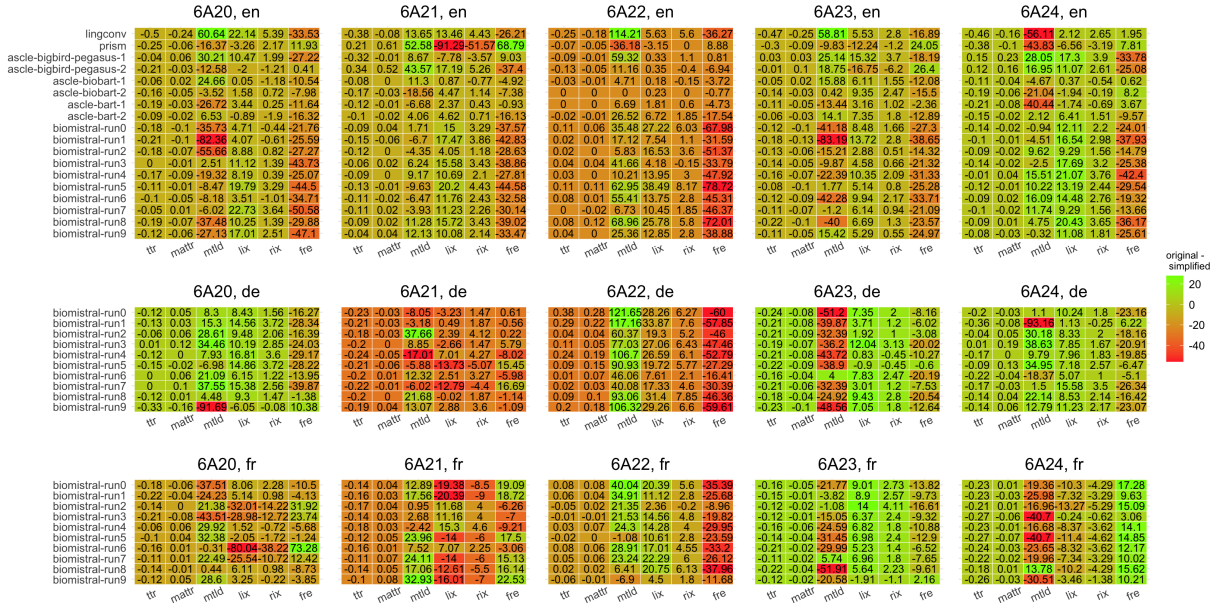


Figure 6: Difference in complexity and readability metrics between the original and simplified texts: across 5 ICD-11 articles (6A20–6A24), 3 languages (English, German, French) and 9 models.

Original text (en)	Original text (de)	Original text (fr)
Schizophrenia is characterised by disturbances in multiple mental modalities, including thinking (e.g., delusions, disorganisation in the form of thought), perception (e.g., hallucinations), self-experience (e.g., the experience that one's feelings, impulses, thoughts, or behaviour are under the control of an external force), cognition (e.g., impaired attention, verbal memory, and social cognition), volition (e.g., loss of motivation), affect (e.g., blunted emotional expression), and behaviour (e.g., behaviour that appears bizarre or purposeless, unpredictable or inappropriate emotional responses that interfere with the organisation of behaviour). Psychomotor disturbances, including catatonia, may be present. Persistent delusions, persistent hallucinations, thought disorder, and experiences of influence, passivity, or control are considered core symptoms. Symptoms must have persisted for at least one month in order for a diagnosis of schizophrenia to be assigned. The symptoms are not a manifestation of another health condition (e.g., a brain tumour) and are not due to the effect of a substance or medication on the central nervous system (e.g., corticosteroids), including withdrawal (e.g., alcohol withdrawal). Schizophrenia, first episode should be used to identify individuals experiencing symptoms that meet the diagnostic requirements for Schizophrenia (including duration) but who have never before experienced an episode during which diagnostic requirements for Schizophrenia were met. Schizophrenia, multiple episodes should be used to identify individuals experiencing symptoms that meet the diagnostic requirements for Schizophrenia and who have also previously experienced episodes during which diagnostic requirements were met, with substantial remission of symptoms between episodes.	Schizophrenie ist durch Störungen in mehreren mentalen Modalitäten gekennzeichnet, einschließlich des Denkens (z. B. Wahnsymptome, Desorganisation in der Form des Denkens), der Wahrnehmung (z. B. Halluzinationen), des Selbsterlebens (z. B. das Erleben, dass die eigenen Gefühle, Impulse, Gedanken oder das Verhalten unter der Kontrolle einer externen Kraft stehen), der Kognition (z. B., Beeinträchtigung der Aufmerksamkeit, des verbalen Gedächtnisses und der sozialen Kognition), des Willens (z. B. Motivationsverlust), des Affekts (z. B. reduzierter Gefühlsausdruck) und des Verhaltens (z. B. Verhalten, das bizarr oder zwecklos erscheint, unvorhersehbare oder unangemessene emotionale Reaktionen, die die Organisation des Verhaltens stören). Psychomotorische Störungen, einschließlich Katatonie, können vorhanden sein. Anhaltende Wahnvorstellungen, anhaltende Halluzinationen, Denkstörungen und Beeinflussungserleben gelten als Kernsymptome. Die Symptome müssen mindestens einen Monat lang bestanden haben, damit die Diagnose Schizophrenie gestellt werden kann. Die Symptome sind nicht Ausdruck eines anderen Gesundheitszustands (z. B. eines Hirntumors) und sind nicht auf die Wirkung einer Substanz oder eines Medikaments auf das zentrale Nervensystem (z. B. Kortikosteroide) zurückzuführen, auch nicht auf einen Entzug (z. B. Alkoholentzug). Schizophrenie, erste Episode sollte verwendet werden, um Personen zu identifizieren, deren Symptome die diagnostischen Anforderungen für Schizophrenie (einschließlich Dauer) erfüllen, die aber noch nie eine Episode erlebt haben, in der die diagnostischen Anforderungen für Schizophrenie erfüllt wurden. Schizophrenie, mehrfache Episoden sollte verwendet werden, um Personen zu identifizieren, die Symptome aufweisen, die die diagnostischen Anforderungen für Schizophrenie erfüllen, und die auch schon früher Episoden erlebt haben, in denen die diagnostischen Anforderungen erfüllt waren, wobei die Symptome zwischen den Episoden deutlich zurückgegangen sind.	La schizophrénie se caractérise par des perturbations de multiples modalités mentales, notamment la réflexion (p. ex. idées délirantes, désorganisation dans la forme de la pensée), la perception (p. ex. hallucinations), l'expérience de soi (p. ex. l'impression que ses propres sentiments, impulsions, pensées ou comportements sont sous le contrôle d'une force externe), la cognition (p. ex. trouble de l'attention, de la mémoire verbale et de la cognition sociale), la volition (p. ex. perte de motivation), l'affect (p. ex. expression émotionnelle émoussée) et le comportement (p. ex. comportement qui semble bizarre ou inutile, réactions émotionnelles imprévisibles ou inappropriées qui perturbent l'organisation du comportement). Des troubles psychomoteurs, y compris une catatonie, peuvent être présents. Les idées délirantes persistantes, hallucinations persistantes, troubles de la pensée et expériences d'influence, de passivité ou de contrôle sont considérés comme des symptômes fondamentaux. Les symptômes doivent avoir persisté pendant au moins un mois pour qu'un diagnostic de schizophrénie soit posé. Les symptômes ne sont pas une manifestation d'une autre affection médicale (p. ex. une tumeur cérébrale) et ne sont pas dus à l'effet d'une substance ou d'un médicament sur le système nerveux central (p. ex. corticostéroïdes), y compris à un sevrage (p. ex. sevrage alcoolique). La schizophrénie, premier épisode doit être utilisée pour identifier les individus qui présentent des symptômes satisfaisant les critères de diagnostic de la schizophrénie (y compris la durée) mais qui n'ont jamais présenté d'épisode précédent durant lequel les critères de diagnostic de la schizophrénie étaient satisfaits. Schizophrénie, épisodes multiples doit être utilisé pour identifier les personnes présentant des symptômes qui répondent aux critères diagnostiques de la schizophrénie et qui ont également déjà connu des épisodes au cours desquels les critères diagnostiques ont été remplis, avec une rémission substantielle des symptômes entre les épisodes.

Original text (en)	Original text (de)	Original text (fr)
Some attenuated symptoms may remain during periods of remission, and remissions may have occurred in response to medication or other treatment. Symptoms fulfilling all definitional requirements of Schizophrenia have been present for almost all of the illness course over a period of at least one year, with periods of subthreshold symptoms being very brief relative to the overall course.	Einige abgeschwächte Symptome können während der Remissionsphasen bestehen bleiben, und Remissionen können als Reaktion auf Medikamente oder andere Behandlungen eingetreten sein. Die Symptome, die alle definitorischen Anforderungen der Schizophrenie erfüllen, waren fast während des gesamten Krankheitsverlaufs über einen Zeitraum von mindestens einem Jahr vorhanden, wobei die Perioden mit unterschwelligen Symptomen im Vergleich zum Gesamtverlauf sehr kurz waren.	Certains symptômes atténués peuvent persister pendant les périodes de rémission, et des rémissions peuvent se produire en réponse à des médicaments ou à un autre traitement. Des symptômes satisfaisant tous les critères de définition de la schizophrénie ont été présents pendant la quasi-totalité de l'évolution de la maladie sur une période d'au moins un an, avec des périodes de symptômes en-dessous du seuil très brèves par rapport à l'évolution globale.

Table 2: Original ICD-11 6A20 description of Schizophrenia (incl. 6A20.0, 6A20.1, 6A20.2) in English, German and French.

Simplified text (en, BioMistral, run 8)	Simplified text (en, BioMistral, run 5)
Schizophrenia is a mental illness that affects many different parts of the mind. It can cause problems with thinking, feeling, and behaving. People with schizophrenia might have strange thoughts or beliefs (delusions), see or hear things that aren't really there (hallucinations), feel like their thoughts or actions are being controlled by someone else, or have trouble remembering things or paying attention. They might also feel less motivated, have trouble controlling their emotions, or behave in ways that seem strange or out of place. Sometimes, people with schizophrenia might become stuck in one position for a long time (catatonia). To be diagnosed with schizophrenia, a person must have had symptoms for at least a month that are not caused by another medical condition or by drugs or alcohol. Schizophrenia is usually divided into three categories: first episode, multiple episodes with remission, and persistent. First episode schizophrenia.	Schizophrenia is a mental illness that affects the way you think, feel, act, and sense things. It can cause problems with your thoughts, like having false beliefs or trouble organizing your thoughts. It can also affect your senses, like hearing or seeing things that aren't really there. You may feel like your emotions are out of control or that you're not in control of your own actions. You might have trouble remembering things or paying attention. You may lose motivation or interest in things. People with schizophrenia may behave in strange or unusual ways, like being overly emotional or having trouble controlling their body movements. They may also have trouble understanding social situations or other people's feelings. Psychomotor disturbances, like being stuck in one position for a long time, may be present. To be diagnosed with schizophrenia, you must have symptoms that have lasted for at least a month and that are not caused by another medical condition or

Table 3: Examples of simplified texts in English, 6A20, run 8 (with high increase in MTLD by 37.48 and increase in FRE by 29.88, ranked as best by 2 human annotators) and run 5 (with lower increase in MTLD by 8.47 and increase in FRE by 44.5, rated as worst by 3 human annotators). The second text shows more repetitive structures (highlighted) reflected by smaller difference in MTLD; the text also has a complex medical term "psychomotor disturbances". The truncated sentences at the end are due to limit in tokens number; they were not manually edited.

Simplified text (de, BioMistral, run 9)	Simplified text (de, BioMistral, run 7)
Schizophrenie ist eine Krankheit, bei der Menschen ihre Denkweise, Wahrnehmung, Gefühle und Verhalten stark beeinträchtigt werden. Diese Beeinträchtigungen können in verschiedenen Formen auftreten, wie Wahnvorstellungen, Halluzinationen, das Erleben, dass ihre eigenen Gefühle, Impulse, Gedanken oder Verhalten unter der Kontrolle einer externen Kraft stehen, Beeinträchtigung ihrer Kognition (z. B. Aufmerksamkeit, Gedächtnis, soziale Fähigkeiten), Motivationsverlust, reduzierter Gefühlsausdruck und bizarr oder zweckloses Verhalten. Psychomotorische Störungen, wie Katatonie, können ebenfalls vorhanden sein. Für die Diagnose Schizophren	Schizophrenie ist eine Krankheit, bei der man sich nicht sicher ist, was man denkt, was man sieht oder hört, wie man sich fühlt, wie man denkt, wie man sich verhält, wie erlebt man Gefühle und wie man sich bewegt. Diese Probleme können sehr lange bestehen und können nicht auf eine andere Krankheit oder auf den Gebrauch von Medikamenten zurückgeführt werden. Diese Krankheit kann erstmals auftreten oder kann schon mehrfach auftreten. Wenn sie erstmals auftreten, wird sie Schizophrenie, erste Episode genannt. Wenn sie mehrfach auftreten, wird sie Schizophrenie, mehrfache Episoden genannt. Die Symptome können während der Krankheit wieder zurückgehen oder verschwinden, aber sie werden nie

Table 4: Examples of simplified texts in German, 6A20, run 9 (with increase in MTLD by 91.69 and decrease in FRE by 10.38) and run 7 (with decrease in MTLD by 37.55 and increase in FRE by 39.87). These are two contrastive examples: 1) complex words and no repetitiveness, dense enumeration, too complex; 2) simpler words but high repetitiveness, too simplified. The truncated sentences at the end are due to limit in tokens number; they were not manually edited.

Simplified text (fr, BioMistral, run 3)	Simplified text (fr, BioMistral, run 4)
La schizophrénie est une maladie mentale où plusieurs types de pensées, perceptions, sentiments et comportements sont perturbés. Par exemple, quelqu'un peut avoir des idées délirantes (comme croire qu'une personne célèbre est en amour avec lui), des hallucinations (comme voir ou entendre quelqu'un qui n'est pas réellement là), une impression que ses propres sentiments, impulsions, pensées ou comportements sont sous le contrôle d'une force externe, un trouble de l'attention, de la mémoire verbale et de la cognition sociale, une perte de motivation, une expression émotionnelle émoussée, un comportement qui semble bizarre ou inutile, des réactions émotionnelles imprévisibles ou inappropriées qui	La schizophrénie est une maladie de l'esprit qui affecte plusieurs facettes de la pensée, de la perception, de l'expérience de soi, de la cognition, de la volonté, de l'affect et du comportement. Les symptômes fondamentaux sont des idées délirantes persistantes, des hallucinations persistantes, des troubles de la pensée et des expériences d'influence, de passivité ou de contrôle. Pour être diagnostiquée, les symptômes doivent persister pendant au moins un mois et ne doivent pas être dus à une autre maladie ou à l'effet d'une substance ou d'un médicament. La schizophrénie est une maladie qui peut avoir plusieurs épisodes, avec des périodes de rémission entre les é

Table 5: Examples of simplified texts in French, 6A20, run 3 (with increase in MTLD by 43.51 and decrease in FRE by 23.74) and run 4 (with decrease in MTLD by 29.92 and increase in FRE by 5.68). These are two contrastive examples: 1) complex words and no repetitiveness, dense enumeration, too complex; 2) a bit simpler words (not much difference with original text) but high repetitiveness, too simplified. The truncated sentences at the end are due to limit in tokens number; they were not manually edited.

BioTopicXplor: A Web Tool for Interactive Exploration of PubMed Literature through Reproducible Topics.

Lana Yeganova lana.yeganova@nih.gov
Donald C. Comeau donald.comeau@nih.gov
Won Kim wonkim@nih.gov
Natalie Xie natxie@nih.gov
Shubo Tian shubo.tian@nih.gov
W. John Wilbur john.wilbur@nih.gov
Zhiyong Lu zhiyong.lu@nih.gov

Division of Intramural Research (DIR), NLM, NIH, Bethesda, MD USA 20894

Abstract

The rapid expansion of biomedical literature presents a major challenge for researchers seeking to organize knowledge, identify emerging scientific trends, and explore large PubMed result sets. Existing literature exploration systems primarily rely on predefined ontologies, user-selected features, or clustering approaches that may produce unstable topic structures across runs. We present BioTopicXplor, a tool for interactive exploration of user-defined PubMed article collections through reproducible topic discovery and hierarchical organization.

BioTopicXplor is powered by ConvexTopics, an optimization-based topic modeling framework that formulates topic discovery as a convex optimization problem with mixture model parameters drawn directly from the data. Unlike traditional clustering approaches that depend on random initialization and may converge to local optima, ConvexTopics guarantees convergence to a global optimum, eliminates the need to predefine the number of topics, and produces stable, fine-grained, and interpretable topic structures across repeated runs on the same corpus.

To support intuitive exploration, BioTopicXplor integrates large language models to generate concise literature-grounded summaries and organize fine-grained subtopics into higher-level thematic groups. The system combines reproducible topic modeling with an interactive human-centric interface that enables users to navigate biomedical literature collections from broad conceptual themes to specific subtopics and associated PubMed articles. BioTopicXplor is freely available at: https://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/topics/dev/create.html.

1 Introduction

The exponential growth of biomedical publications presents a critical challenge for researchers seeking to organize, interpret, and discover knowledge or emerging scientific trends. PubMed (https://pubmed.ncbi.nlm.nih.gov/) remains the main gateway to biomedical literature and provides access to relevant documents given a query. However, it does not offer support for understanding the conceptual structure of a collection.

Several companion tools enable alternative options for exploring biomedical literature. Anne O'Tate (Engwall, 2017) enables users to analyze PubMed search results by ranking and grouping articles based on important words in titles and abstracts. It also allows clustering by MeSH term–based topics, author names, affiliations, journal names, and publication year. BioTextQuest v2.0 (Theodosiou et al., 2024), extends this paradigm by offering interactive clustering based on user-selected biomedical terms. Ontology-based interfaces such as GoPubMed (Doms and Schroeder, 2005) organize PubMed records within structured vocabularies, including Gene Ontology and medical Subject Headings (MeSH), facilitating exploration but remaining constrained to predefined ontological categories rather than enabling unsupervised topic discovery. While these tools facilitate exploration, they rely on predefined features and user-selected terms, or curated ontologies to organize the literature. As a result, they may be limited in their ability to capture the data-driven structure of a document collection or to reveal fine-grained, emergent topics in an unsupervised way.

To address these limitations, we present BioTopicXplor, a tool for interactive exploration of user-defined PubMed article collections. Given a PubMed query, BioTopicXplor performs on-demand clustering/topic discovery and organizes the resulting topics into an intuitive representation that supports exploration of literature collections both through navigating the topics and querying within the topics.

BioTopicXplor is powered by ConvexTopics

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 475–480
July 3-4, 2026 ©2026 Association for Computational Linguistics

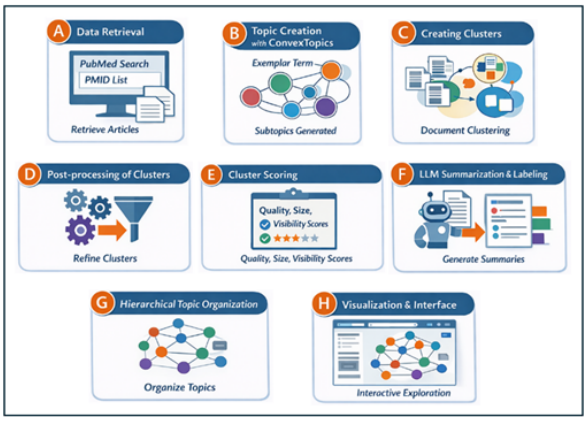


Figure 1: BioTopicXplor architecture.

(Yeganova et al., 2026), an optimization-based topic modeling framework that is formulated as a convex optimization problem with exemplars drawn directly from the data (Lashkari and Golland, 2007). We chose this approach to avoid limitations of traditional clustering methods such as K-means (Ja, 1979) or expectation–maximization for mixture models (Blei et al., 2003). These suffer from sensitivity to initialization and convergence to local optima, limiting reproducibility and scalability on large corpora. As a result, these methods generally produce different document or term groupings across runs.

ConvexTopics guarantees convergence to a global optimum, which means it yields the same set of topics across repeated runs on the same document collection. Additionally, it eliminates dependence on random initialization, and the need to predefine the number of topics. The algorithm produces a large number of fine-grained topics, referred to as subtopics.

To support interpretation and exploration of these subtopics, BioTopicXplor integrates large language models for two key tasks: (1) to generate concise summaries for each subtopic, and (2) to construct a second layer that organizes subtopics into higher-level groups, referred to as major topics. This hierarchical structure enables users to navigate from broad thematic areas to more specific ones. In addition, the tool provides a search functionality that allows users to query and explore subtopics interactively. We demonstrate the utility of the system through application to anti-aging literature, illustrating its ability to reveal meaningful thematic structure leading to knowledge discovery.

2 System Overview

Given a user-specified PubMed query, BioTopicXplor retrieves the corresponding articles, applies the ConvexTopics framework to generate a set of subtopics, and subsequently organizes them into higher-level, more general topics. The overall pipeline (illustrated in Figure 1) integrates data retrieval, text processing, optimization-based topic modeling, document clustering, and visualization components, as well as large language models to generate summaries for each subtopic, allowing users to seamlessly move from raw literature collections to structured, interpretable topics.

As LLMs become increasingly integrated into biomedical text mining, their ability to generate user-friendly summaries is highly valuable. However, they remain prone to hallucination. ConvexTopics provides a stable backbone that constrains the use of LLMs, ensuring that summaries are grounded in algorithmically derived topics. We believe this controlled integration provides the benefit of human-readable summaries while minimizing hallucinations by keeping the summaries tightly anchored to the topic terms and relevant documents computed by ConvexTopics.

Data Retrieval. BioTopicXplor begins by retrieving a set of PubMed articles based on a user-specified query. By default, the retrievals are limited to the last five years. The system collects titles and abstracts for downstream processing.

Creating Subtopics with ConvexTopics The subtopics are computed using the ConvexTopics algorithm (Yeganova et al., 2026), which reformulates clustering as a convex optimization problem by restricting potential cluster centers to exemplars drawn directly from the data. This transforms what is ordinarily a combinatorial search into a tractable convex objective that is guaranteed to converge to a globally optimal solution (Lashkari and Golland, 2007). Applied to biomedical text, the algorithm operates over a vocabulary of statistically enriched single terms and noun phrases, as many biomedical concepts are expressed as multi-word expressions (Yeganova et al., 2009, 2011). Term similarity is measured via the Dice coefficient (Manning, 2008). The number of resulting subtopics is determined automatically during computation, yielding fine-grained, reproducible topic structures without manual tuning. Each subtopic is defined by an exemplar and a set of representative terms that are used to rank documents by relevance.

Creating Document Clusters. For each subtopic created above, we use its topic terms to compute the list of the most relevant PubMed articles. These documents are selected based on their relevance to the topic terms and provide a cluster of documents associated with each topic. Clicking on a PubMed ID or article title leads to the article view in PubMed, where one can read, download, or follow links to full-text, if available.

Post-processing of Clusters. Following initial clustering, a post-processing step is applied. Redundant clusters and clusters with substantial overlap (greater than 50% shared documents) are removed to ensure a diverse and non-redundant set of subtopics.

Cluster Scoring. Each topic/cluster is assigned three scores based on the following criteria: (i) an intrinsic cluster quality score, (ii) the size of the cluster measured by the number of associated documents, and (iii) a visibility score that reflects the prominence of the ten top scoring topic terms in the ten top scoring documents. Each of these scores can be used to rank subtopics for presentation.

Summarization and Labeling. BioTopicXplor employs large language models to generate concise, literature-grounded summaries and descriptive titles for each subtopic. The summaries are generated based on the top ten documents associated with each cluster using GPT-5.1. Each generated summary sentence is linked to PMID(s) that were used for extracting the underlying evidence. The title is generated to best describe the summary.

Hierarchical Topic Organization. To organize the large number of subtopics for presentation, they are grouped into higher-level topics (major topics). This step leverages large language models to group semantically related subtopics and construct a hierarchical representation.

Visualization and Interface. The resulting higher-level topics are presented through an interactive interface that allows users to explore clusters, view summaries, and access associated PubMed articles. The visualization is designed to provide an overview of the conceptual structure of literature.

Implementation Details. The user-facing web interface was developed in-house using standard HTML, CSS, and JavaScript. The front end is compatible with modern desktop browsers and iPads. Although not specifically optimized for phone screens, the server is functional on mobile devices. User requests are submitted to the server, which was also developed in-house using C++ and Python. Both the front end and backend components run on a Linux-based system. The system runs on a shared machine equipped with a 12-core Xeon(R) Gold CPU operating at up to 2.30 GHz and 196 GB of RAM. Most of the processing time is spent generating cluster titles and summaries using an OpenAI GPT model, currently GPT-5.1.

3 Usage

Users initiate an analysis by submitting a PubMed query, after which the system automatically retrieves the corresponding articles and executes the full processing pipeline. The results are provided through a uniquely generated web link. Multiple analyses can be conducted concurrently, and previously generated instances remain available for subsequent exploration.

View Major Topics. On the main page, users are presented with Major Topics derived for the dataset. These serve as starting points for exploration. Selecting a Major Topic leads to corresponding Subtopics, which can be ranked according to multiple scoring criteria.

Explore Subtopics. Each subtopic is associated with a dedicated page that includes a concise summary and title generated from the ten most relevant documents, a ranked list of associated PubMed articles with links for direct access, and a set of topic terms defining the topic.

Explore the data through Search. In addition to hierarchical navigation, the system supports keyword-based search across topics. Search can be performed globally across the entire collection or restricted to a specific major topic, enabling targeted exploration of relevant subtopics.

Use Case: Exploring Anti-Aging literature in PubMed. The growing interest in longevity and anti-aging has led to a proliferation of resources promoting various interventions, not all of which are supported by rigorous scientific evidence. This underscores the need for systematic analysis of biomedical literature to distinguish evidence-based claims. Using anti-aging research as a representative use case, (Yeganova et al., 2026) demonstrates how BioTopicXplor, applied to 12,000 PubMed articles, captures the research landscape through nuanced subtopics with summaries, and hierarchical organization.

Figure 2 illustrates the framework. The left panel shows the main page, where users are presented with a list of major topics. The right panel shows

a detailed topic page. Demonstrated in Figure 2 is a topic page produced by choosing the "Gut Microbiota and aging" major topic and selecting the "Beneficial Bacteria and Natural Compounds Combat Aging" subtopic.

Using this analysis, (Yeganova et al., 2026) observed that although NAD supplements are widely available, their safety remains insufficiently supported by evidence. Key uncertainties include pharmacokinetics and pharmacodynamics, particularly bioavailability, metabolism, and tissue specificity (Poljšak et al., 2022). Moreover, the lack of long-term safety studies and limited clinical trials leaves optimal dosing unclear, with some studies suggesting a potential risk of kidney inflammation in older populations (Song et al., 2023; Nadeeshani et al., 2022).

4 Evaluation

Many factors can be considered for evaluating BioTopicXplor as an end-to-end system, examining not only the quality of the topics produced by ConvexTopics but also the interpretability and utility of the resulting framework.

The quality of Subtopics. The ConvexTopics algorithm has been evaluated on both general-purpose NLP benchmarks and biomedical text corpora in (Yeganova et al., 2026). The annotated baselines include 20-Newsgroups (`http://people.csail.mit.edu/jrennie/20Newsgroups`) and Reuters-21578 (Lewis et al., 2004), while the biomedical datasets consist of PubMed articles on Anti-Aging, Diabetes Mellitus, and Age-Related Macular Degeneration. ConvexTopics was compared with state-of-the-art clustering and topic modeling approaches, including K-means (Ja, 1979; Lloyd, 1982), LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022) and was shown to perform favorably.

The quality of Major Topics. To facilitate navigation, BioTopicXplor organizes subtopics into major topics using large language models. We evaluated this hierarchical structure through manual inspection. The major topics were observed to correspond to meaningful thematic groupings, providing a coherent high-level view of the literature while preserving access to detailed subtopics. While some variations in major topics may occur accross the runs, the resulting groupings remain meaningful.

Timing. Depending on the size of the dataset, processing may take up to several hours. For example, analysis of 1,099 documents retrieved using the query “familial Mediterranean fever” (limited to 5 most recent years) completed in 65 minutes, while the same query without time restriction produced 5,731 documents and required 87 minutes. The relatively small increase in processing time between 1,099 and 5,731 documents is explained by the fact that overall runtime is dominated not by ConvexTopics clustering itself, but by downstream LLM-based summarization and hierarchical organization steps. These stages depend primarily on the number of generated subtopics and summaries. We are currently working on enabling users to access preliminary results generated by ConvexTopics, allowing them to access the subtopics before the full set of summaries and hierarchical organization is completed.

5 Discussion and Conclusions

BioTopicXplor provides a scalable, reproducible framework for organizing and exploring biomedical literature. It enables exploration not possible with PubMed alone, and is particularly valuable for researchers entering a new area or seeking deeper understanding of a topic. It may also benefit a broader audience by providing a human-interpretable overview of terminology-heavy PubMed literature. We demonstrate its effectiveness through application to the rapidly growing field of anti-aging. Across domains, BioTopicXplor reveals meaningful substructures and maintains consistent topic organization.

Several limitations should be noted. ConvexTopics tends to produce a large number of subtopics, requiring additional post-processing and hierarchical organization to improve interpretability. Additionally, as a web-based system, BioTopicXplor is subject to computational and latency constraints when processing very large document collections in real time. A substantial portion of the processing time is spent generating topic summaries and titles using large language models. These titles are important for comprehension, and play a role in organizing subtopics into higher-level topics. We find that the improved interpretability and resulting organizational structure provided to the user justify the increased processing time.

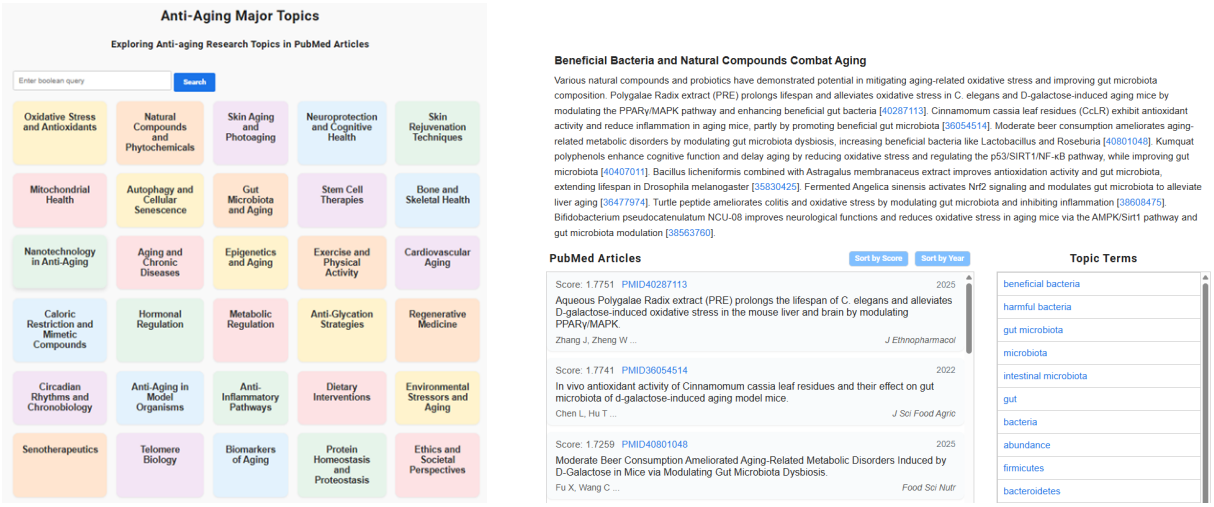


Figure 2: The visualization of the BioTopicXplor applied to the Anti-Aging literature in PubMed. Left panel shows the high-level topics displayed on the main page. Clicking one of these major topics leads to a list of subtopics relevant to the major topic. Clicking on a subtopic leads to a topic page containing topic terms, related pmids, topic summary and title, presented on the right panel. The graphics are adapted from the earlier ConvexTopics paper.

6 Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Andreas Doms and Michael Schroeder. 2005. Gopubmed: exploring pubmed with the gene ontology. *Nucleic acids research*, 33(suppl_2):W783–W786.

Keith D Engwall. 2017. Anne o'tate. *Journal of the Medical Library Association: JMLA*, 105(2):200.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hartigan Ja. 1979. A k-means clustering algorithm. *JR Stat. Soc. Ser. C-Appl. Stat.*, 28:100–108.

Danial Lashkari and Polina Golland. 2007. Convex clustering with exemplar-based models. *Advances in neural information processing systems*, 20.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Harshani Nadeeshani, Jinyao Li, Tianlei Ying, Baohong Zhang, and Jun Lu. 2022. Nicotinamide mononucleotide (nmn) as an anti-aging health product–promises and safety concerns. *Journal of advanced research*, 37:267–278.

Borut Poljšak, Vito Kovač, and Irina Milisav. 2022. Current uncertainties and future challenges regarding nad+ boosting strategies. *Antioxidants*, 11(9):1637.

Qin Song, Xiaofeng Zhou, Kexin Xu, Sishi Liu, Xinqiang Zhu, and Jun Yang. 2023. The safety and antiaging effects of nicotinamide mononucleotide in human clinical trials: an update. *Advances in nutrition*, 14(6):1416–1435.

Theodosios Theodosiou, Konstantinos Vrettos, Ismini Baltsavia, Fotis Baltoumas, Nikolas Papanikolaou, Andreas N Antonakis, Dimitrios Mossialos, Christos A Ouzounis, Vasilis J Promponas, Makrina Karaglani, and 1 others. 2024. Biotextquest v2. 0: an evolved tool for biomedical literature mining and concept discovery. *Computational and Structural Biotechnology Journal*, 23:3247–3253.

Lana Yeganova, Donald C Comeau, Won Kim, and W John Wilbur. 2009. How to interpret pubmed queries and why it matters. *Journal of the American Society for Information Science and Technology*, 60(2):264–274.

Lana Yeganova, Donald C Comeau, and W John Wilbur. 2011. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC bioinformatics*, 12(Suppl 3):S6.

Lana E. Yeganova, Won G. Kim, Shubo Tian, Natalie Xie, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2026. Exploring anti-aging literature via convextopics and large language models. *AMIA Amplify Informatics Conference*.

Reading Between the Lines: Toward Translating Verbose Patient-authored Messages into Clinician-Formulated Questions

Sarvesh Soni, Madeline Bittner, Dina Demner-Fushman
Division of Intramural Research
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
{sarvesh.soni, madeline.bittner}@nih.gov, ddemner@mail.nih.gov

Abstract

Patient portal messages often embed clinical questions inside long, emotionally nuanced narratives, requiring clinicians to infer the underlying information need. We study the task of rewriting verbose patient-authored narratives into concise, clinician-interpreted questions framed as if querying an electronic health record (EHR) system. We evaluate a lightweight LLM-based rewrite pipeline that constrains outputs to 10–15 words and uses rule-based validation with regeneration. We test the approach on 140 distinct patient questions drawn from the ArchEHR-QA dataset and shared task. Each system output is double-annotated by two annotators for quality (Good/Ok/Bad) and error types (Generic, Malformed, Tangential, Hallucination). Results show that while models follow output constraints, they often produce overly generic or tangential questions, and occasional hallucinations introduce unsupported clinical details. Across both clinician-question and patient-narrative comparison settings, automatic metrics show substantial overlap across human quality labels; in pairwise meta-evaluation, BERTScore is the strongest proxy for human preferences. We release our code and annotations to support future work[1].

1 Introduction

Patients frequently contact healthcare teams via patient portals with questions about tests, diagnoses, and care plans (Tai-Seale et al., 2017). In practice, these questions are often embedded in lengthy narratives with extraneous detail, uncertainty, and emotion (Table 1), forcing clinicians to "read between the lines" to determine what the patient is actually asking (Dymek et al., 2021; Yan et al., 2021). Translating patient-authored narratives into concise clinician-interpreted questions is key for downstream systems: it clarifies the clinical information needs, enables targeted retrieval of supporting electronic health record (EHR) evidence, and reduces the risk that automated assistance addresses the wrong intent. We view this reformulation as an auxiliary representation that complements, rather than replaces, the original patient message.

Patient Question: Hi I ... started shaking ... woken up ... confused ... been told that there is a shadow on my brain ...
Clinician-interpreted Question: What caused the patient's confusion and short-term memory loss?
Output #1 (**Quality:** *Ok*; **Error:** *Generic*): What are possible causes of a brain shadow after seizure activity?
Output #2 (**Quality:** *Bad*; **Error:** *Hallucination*): What is the likely cause of the new cerebral lesion on imaging?

Table 1: Abbreviated example of question reformulation and human evaluation; full version in Appendix Table 5.

Prior work on medical question summarization targets generic consumer health queries answerable from general medical knowledge (Yadav et al., 2022). In contrast, answers to patients' EHR-related questions depend on hospitalization records. Therefore, reformulation must (i) remain faithful to patient intent, (ii) support patient-record retrieval, and (iii) avoid medical details not supported by the message. Moreover, existing methods rarely model patient-to-clinician perspective shift. Prior natural language processing work on portal messages has emphasized triage or generic response generation (Anderson et al., 2025; Biro et al., 2025); however, reformulating EHR-grounded patient narratives into clinician-style questions is underexplored. We make three contributions:

1. We present a simple large language model (LLM)-based system with constrained output, rule-based validation, and regeneration for patient-to-clinician question reformulation.

2. We conduct double-annotated human evaluation of LLM outputs, rating quality (*good*,

[1] https://github.com/soni-sarvesh/patient-question-reformulation

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 481–489
July 3-4, 2026 ©2026 Association for Computational Linguistics

ok, or *bad*) and categorizing errors (*generic*, *malformed*, *tangential*, *hallucination*).

3. We analyze automatic metrics in two comparison settings (against the reference clinician question and the patient-authored narrative) and meta-evaluate which metrics best track human preferences for model comparison.

2 Related Work

Medical question summarization. Generating clinician-interpreted questions from patient narratives is question summarization, which is widely studied for consumer health questions answerable from general medical knowledge rather than patient-specific records (Ben Abacha et al., 2021).

EHR question answering (QA) and information retrieval (IR). Most clinical QA and IR work focuses on clinician-authored questions over EHR content (Bardhan et al., 2024; Sivarajkumar et al., 2024). Differently, ArchEHR-QA (Soni and Demner-Fushman, 2026a) is a publicly available resource that pairs patient-authored narratives with clinician-interpreted questions and annotated EHR evidence for patient-specific QA. We build on this line of work by isolating the upstream *reformulation* step, focusing on translating patient language into a clinician-formulated query suitable for downstream retrieval and grounding.

Patient portal messaging. Prior work on patient portal messages has emphasized message classification and triage to reduce clinician burden and improve routing (Tafti et al., 2019; Ren et al., 2023; Anderson et al., 2025). More recently, large language models (LLMs) have been integrated for drafting responses to patient messages, raising both efficiency opportunities and safety concerns (Chen et al., 2024; Garcia et al., 2024; Biro et al., 2025). Clinician question reformulation is complementary: by surfacing the inferred clinical information need, it can support safer downstream retrieval of patient-specific evidence and grounded response drafting.

3 Methods

3.1 Data

We use patient-authored questions and corresponding clinician-interpreted questions from ArchEHR-QA (Soni and Demner-Fushman, 2026a), which links questions from public health forums about recent hospitalizations to de-identified clinical notes from MIMIC (Johnson et al., 2016, 2023). To preserve clinical fidelity, the clinical notes are unedited; the patient questions are minimally edited to match surface details in the associated note (e.g., pronoun changes such as "he" to "she"). In total, we evaluate on 140 distinct patient questions. This set combines the 108 unique questions in the original ArchEHR-QA release (134 question–note instances before deduplication) with 32 additional questions from the ArchEHR-QA 2026 Shared Task data (Soni and Demner-Fushman, 2026b).

3.2 Model

We use a simple LLM-based generation approach with rule-based validation. Given a patient-authored narrative, an instruction-tuned LLM is prompted to produce a concise clinician-interpreted question that captures inferred clinical intent and is suitable for downstream EHR-oriented use. The prompt constrains the output to a single 10–15-word clinician question to promote uniformity (Table 4). Validation uses deterministic string processing to extract the clinician question from the prompted markdown format and verify that exactly one non-empty question is present. If extraction fails (i.e., the generated output does not conform to the prompted format), the same prompt is reused for up to five attempts; the first generation that passes the format check is retained. In our runs, every retained output passed validation within five attempts. This rule-based check only enforces output format, not the 10–15-word length; the length constraint is specified in the prompt and influences generation, but is not strictly enforced post-hoc. Beyond this, we do not use retrieval, external medical knowledge, or hand-crafted rewriting rules. We evaluate two open-weight LLMs with different parameter scales: Llama 3.1 8B and 3.3 70B (Grattafiori et al., 2024); decoding and inference settings are reported in Appendix A.1.

3.3 Evaluation

Manual evaluation. Three annotators with clinical/biomedical expertise (an MD, a clinical informaticist, and a biomedical researcher) evaluated model outputs for *Quality* (*Good*: faithful and specific; *Ok*: usable but imperfect; *Bad*: unusable) and assigned *Error* labels as needed. We used four non-mutually-exclusive error categories: *Generic* (too general), *Malformed* (ill-formed), *Tangential* (not focused on the core patient intent), and *Hallucination* (unsupported details). Annotators could view

3-class *Good / Ok / Bad*		Binary {*Good, Ok*} / *Bad*	
P_o	κ	P_o	κ
36.8%	0.162	71.1%	0.234

Table 2: Inter-rater agreement on *Quality* annotations for 3-class (*Good*/*Ok*/*Bad*) and a binary recoding (*Acceptable*/*Bad*). Exact match P_o and Cohen's κ are reported.

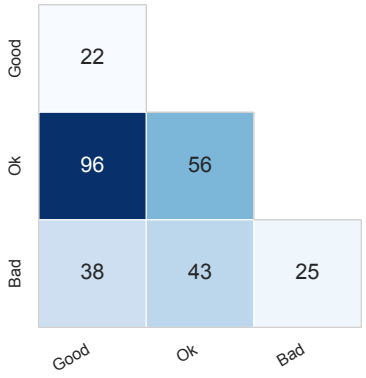


Figure 1: Pairwise agreement/disagreement matrix for *Quality* ratings between annotators ($N = 280$). Diagonal cells show agreement counts; off-diagonal cells show disagreement counts for each label pair.

the reference clinician question, but overlap with the reference was not a rating criterion, since multiple reformulations could be valid; annotators were blinded to model identity. All 280 outputs (140 cases×2 models) were double-annotated. For the first 50 cases (100 outputs), annotators discussed disagreements to calibrate guideline interpretation, but did not revise labels for subjective disagreements; the remaining cases were annotated independently, and we retain both labels without adjudicating subjective disagreements.

Automated evaluation. We compute ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), Medcon (Yim et al., 2023), and AlignScore (Zha et al., 2023) between each generated question and two comparison texts: the reference clinician question and the patient-authored narrative. These metrics capture lexical overlap (ROUGE), embedding similarity (BERTScore), clinical concept overlap (Medcon), and factual consistency (AlignScore).

4 Results and Discussion

The average lengths of the patient narratives and reference clinician-interpreted questions are 100.4 and 11.9 words, respectively. The Llama 8B and

Set-level (multi-label) agreement	
Metric	**Score**
Mean Jaccard similarity	0.283
Exact match rate (%)	19.3

Per-error-type agreement		
Label	P_o(%)	κ
Generic	45.7	−0.086
Malformed	93.2	0.629
Tangential	60.7	0.070
Hallucination	84.6	0.230

Table 3: Inter-rater agreement for *Error* annotations. Set-level agreement for multi-label error sets uses mean Jaccard similarity and exact match; per-error-type agreement reports percent agreement P_o and Cohen's κ.

70B models generated questions averaging 15.5 and 11.5 words, respectively. Since the 10–15-word constraint is conveyed via the prompt but not enforced by validation, these averages reflect how closely each model follows the prompted length instruction. The 70B model adheres tightly to the prompted range: 95% of its outputs fall within 10–15 words (with the remaining 5% slightly under 10). The 8B model follows the lower bound reliably (100% of outputs are at least 10 words) but frequently exceeds the upper bound, with only 48.6% of outputs in the 10–15 range.

Inter-rater Agreement. On *Quality*, annotators agreed on 36.8% of instances, increasing up to 71.1% with binary recoding (Table 2). This is consistent with a subjective boundary between *Good* and *Ok* and with the difficulty of judging adequacy; Figure 1 shows the distribution of disagreements, and Cohen's κ indicates slight agreement.

For *Error* labels, set-level agreement is moderate with the mean Jaccard similarity of 0.283, while the exact match is 19.3% (Table 3). Per-label agreement varies widely: *Malformed* errors are easiest to identify with a percent agreement (P_o) of 93.2%, whereas *Generic* is hardest ($P_o = 45.7\%$), likely because judgments about whether a reformulation is overly broad or insufficiently specific are more subjective than judgments about surface forms.

Inter-rater variability is also common in clinical interpretation itself, where specialists may legitimately disagree (Novack et al., 2006). Related open-ended biomedical annotation tasks, including Medical Subject Headings (MeSH) indexing and evaluation of generated clinical notes, also report modest inter-rater agreement (Fernandez-Llimos, 2025; Moramarco et al., 2022). We therefore pre-

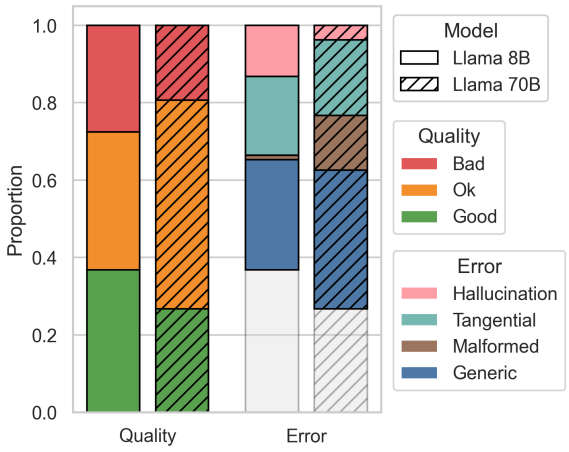


Figure 2: Human judgments of *Quality* and *Error* for questions generated by Llama 8B and 70B models.

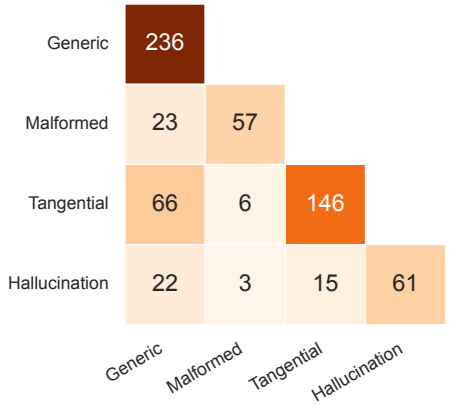


Figure 3: *Error*-label co-occurrence matrix computed over all annotator-output instances (no adjudication). Diagonal: count of each error label; off-diagonal: within-instance co-selections of error pairs.

serve these annotations as complementary annotator judgments rather than as a single gold label.

Human ratings. Figure 2 summarizes *Quality* and *Error* judgments for both models. Overall, the 70B model produces fewer *Bad* outputs, but many generations remain *Ok*, reflecting that partially correct intent capture may still be too generic or miss patient-specific nuance. Both models exhibit substantial *Generic* and *Tangential* errors; the 8B model shows a higher fraction of *Hallucination*-labeled errors, while the 70B model shows relatively more *Malformed*-labeled errors.

Error co-occurrence. Figure 3 characterizes error frequency and co-occurrence across annotated *Error* label instances (error categories are non-mutually exclusive). *Generic* is the dominant failure mode (236 instances), followed by *Tangential* (146), with *Hallucination* (61) and *Malformed* (57) occurring less frequently. *Generic* frequently co-occurs with *Tangential* (66 instances), suggesting that a model latches onto a peripheral detail and produces a broad question about it rather than the most clinically relevant information need.

Hallucinations are less common but clinically important. As shown in Table 1, hallucinations can introduce unsupported details (e.g., assuming a "lesion" when the patient only mentions a "shadow"), which would misdirect downstream EHR retrieval and response drafting. Hallucination co-occurs more often with *Generic* (22) or *Tangential* (15) than with *Malformed* (3), suggesting that hallucinations are often fluent and well-formed, making them harder to detect using format checks alone.

Automatic metrics and human quality judgments. ROUGE, BERTScore, Medcon, and AlignScore stratified by human *Quality* labels are shown in Appendix Figures 4 and 5. When computed against the reference clinician question, automatic metrics align only weakly with human *Quality* ratings. BERTScore shows the clearest upward shift for both 8B and 70B models (higher median for *Good* than *Bad*/*Ok*). ROUGE and AlignScore show heavy overlap across *Bad*/*Ok*/*Good*, and the median scores for *Good* are not consistently higher than those for *Ok* or *Bad*. Medcon is largely zero-dominated with high variance and overlap across quality levels, consistent with sparse concept overlap when only a single reference clinician question is available and when acceptable reformulations may use different clinically meaningful terms. Thus, automated metrics may flag very low-scoring outliers but do not reliably distinguish reformulations with different *Quality* labels and may miss clinically consequential errors (e.g., fluent hallucinations), reinforcing the need for expert evaluation and richer multi-reference benchmarks. Using the patient-authored narrative instead yields the same broad conclusion: metric score distributions still overlap substantially across quality labels. However, the clinician-question setting appears somewhat more informative overall, most clearly for BERTScore, while differences for the other metrics are smaller or less consistent.

Automatic metrics and human pairwise preferences. We meta-evaluate automatic metrics against human pairwise preferences, following

prior work (Freitag et al., 2024), on the subset of cases where the two systems received different human quality ratings (Appendix Table 6). When computed against the reference clinician question, BERTScore achieves the highest overall pairwise accuracy (PA; 61.0%) and is the best metric in each quality-pair bucket, whereas ROUGE-L and AlignScore remain below 50% PA overall. When computed against the patient narrative, BERTScore still ranks first (52.0%), but the gap to ROUGE-L and AlignScore narrows (both 48.0%). Medcon remains weaker, though its PA increases from 26.8% with clinician questions to 40.7% with patient narratives. Overall, BERTScore is the strongest proxy for human preferences in our setting.

5 Conclusion

In a simple LLM-based constrained generation approach to translating verbose patient-authored narratives into concise clinician-interpreted questions, with outputs assessed through double-annotated human evaluation, we found that fluent rewrites are often *generic* or *tangential*, and occasional *hallucinations* introduce unsupported clinical details. Automatic metrics do not cleanly separate more useful reformulations from less useful ones; in pairwise meta-evaluation, BERTScore provides the most informative signal for relative model comparison among the metrics we tested.

Limitations

First, clinician interpretation of a patient narrative is inherently subjective, and we do not adjudicate disagreements; instead, we retain multi-annotator labels. While this better reflects variation in clinician reasoning, it complicates learning and evaluation with a single "gold" label. Second, each instance currently provides only one reference clinician-interpreted question, even though multiple rewrites may be equally acceptable. This is mitigated by the judgments on the generated questions that could now be used as paraphrases in future evaluations. Third, we evaluate two open-weight models from a single family (Meta Llama 3.1 8B and 3.3 70B). This is a deliberate scope choice: the paper's contribution is a characterization of reformulation error modes and a meta-evaluation of automatic metrics, rather than a leaderboard-style model comparison. The specific error rates and metric rankings may not transfer to other open-weight, proprietary, or clinically adapted LLMs. Fourth, the 10–15-word output constraint was chosen to promote uniformity and to discourage models from echoing patient narrative verbatim, but it can force compression of multi-part patient concerns into a single question. The reference clinician questions in ArchEHR-QA average 11.9 words, and themselves often collapse compound concerns; we did not separately quantify how often important nuance is lost by the constraint. Relaxing or removing the length cap, or allowing multiple sub-questions, is worth exploring in future work. Fifth, the evaluation set contains 140 patient cases, so small between-model differences should not be over-interpreted. Sixth, we report one final valid generation per model–instance pair under a fixed prompting and decoding setup. The pipeline is intentionally simple and is not tuned to maximize task performance through extensive prompt engineering, retrieval augmentation, or model adaptation. Stronger systems or averaged multi-sample evaluations could change the absolute performance levels and possibly the relative behavior of the automatic metrics. Finally, our evaluation isolates the reformulation step and does not measure downstream effects such as evidence retrieval or answer grounding; assessing these is a natural next step.

Ethical considerations

This work uses a publicly available, de-identified dataset and releases additional annotations for research. Nevertheless, the task involves clinical language and could be misused if deployed without oversight. In particular, hallucinated clinician questions could misdirect evidence retrieval or clinical reasoning. We position our work as a benchmark and analysis of failure modes, and any real-world deployment should include clinician-in-the-loop review, monitoring for hallucinations, and safeguards against inappropriate clinical decision-making.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Blake J. Anderson, Muhammad Zia ul Haq, Yuanda Zhu, Andrew Hornback, Alison D. Cowan, Michelle Mott, Bradley Gallaher, and Arash Harzand. 2025. Development and Evaluation of a Model to Manage Patient Portal Messages. *NEJM AI*, 2(3):AIoa2400354.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question Answering for Electronic Health Records: Scoping Review of Datasets and Models. *Journal of Medical Internet Research*, 26(1):e53636.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Joshua M. Biro, Jessica L. Handley, J. Malcolm McCurry, Adam Visconti, Jeffrey Weinfeld, J. Gregory Trafton, and Raj M. Ratwani. 2025. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digital Medicine*, 8(1):1–6.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 0(0).

Christine Dymek, Bryan Kim, Genevieve B Melton, Thomas H Payne, Hardeep Singh, and Chun-Ju Hsiao. 2021. Building the evidence-base to reduce electronic health record–related clinician burden. *Journal of the American Medical Informatics Association*, 28(5):1057–1061.

Fernando Fernandez-Llimos. 2025. Consistency of Medical Subject Headings assignment: A test-retest reliability analysis. *Research in Social and Administrative Pharmacy*, 21(10):784–789.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Patricia Garcia, Stephen P. Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, and Christopher Sharp. 2024. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Network Open*, 7(3):e243201.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Victor Novack, Lone S. Avnon, Alexander Smolyakov, Rachel Barnea, Alan Jotkowitz, and Francisc Schlaeffer. 2006. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. *European Journal of Internal Medicine*, 17(1):43–47.

Yang Ren, Dezhi Wu, Aditya Khurana, George Mastorakos, Sunyang Fu, Nansu Zong, Jungwei Fan, Hongfang Liu, and Ming Huang. 2023. Classification of Patient Portal Messages with BERT-based Language Models. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 176–182.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical Information Retrieval: A Literature Review. *Journal of Healthcare Informatics Research*, 8(2):313–352.

Sarvesh Soni and Dina Demner-Fushman. 2026a. A Dataset for Addressing Patient's Information Needs

related to Clinical Course of Hospitalization. *Scientific Data*, 13(1):523.

Sarvesh Soni and Dina Demner-Fushman. 2026b. Overview of the ArchEHR-QA 2026 shared task on grounded question answering from electronic health records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (Cl4health)*, Palma, Mallorca (Spain). ELRA.

Ahmad P. Tafti, Sunyang Fu, Aditya Khurana, George M. Mastorakos, Kenneth G. Poole, Stephen J. Traub, James A. Yiannias, and Hongfang Liu. 2019. Artificial intelligence to organize patient portal messages: A journey from an ensemble deep learning text classification to rule-based named entity recognition. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1380–1387.

Ming Tai-Seale, Cliff W. Olson, Jinnan Li, Albert S. Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S. Luft. 2017. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. *Health Affairs*, 36(4):655–662.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040.

Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. 2021. Exploring the relationship between electronic health records and provider burnout: A systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Implementation Details

This section reports decoding settings and the complete generation prompt.

You are provided with a patient question under `[PATIENT NARRATIVE]`.

`[PATIENT NARRATIVE]:`

`{...}`

Your task is to generate a clear and concise clinician question that a clinician would interpret from the patient narrative. Assume that the clinician is asking this question to a smart electronic health record system to help themselves in formulating a response to the patient. Restrict the length of the clinician question between 10 to 15 words.

Format instructions: Output your response in a markdown format, containing only the clinician question. For example:

```clinician-question
This is a clinician-version of the patient question.
```

Table 4: Complete prompt used to automatically generate the clinician-interpreted questions.

A.1 Hyperparameters

We use fixed decoding parameters for all runs (HuggingFace Transformers text-generation pipeline): do_sample=True, temperature=0.9, top-p=0.6, top-k=50, repetition_penalty=1.05, and max_new_tokens=512; we do not perform hyperparameter search. For each model and patient narrative, we generate a single final clinician question (with up to five retries only when the output fails format/length validation); we do not sample multiple valid generations or average across random seeds.

A.2 Prompt

Table 4 gives the complete prompt used to generate clinician-interpreted questions.

B Supplementary Example

Table 5 provides the full version of the abbreviated example in Table 1.

C Supplementary Analyses

This section provides supplementary figures, tables, and analyses that complement the results in the main text.

C.1 Quality-rating disagreements

Figure 1 shows that disagreements were mostly between adjacent labels (*Good* vs. *Ok*: 96; *Bad* vs. *Ok*: 43) rather than the extremes (*Good* vs. *Bad*: 38), supporting the use of multi-annotator labels rather than a single adjudicated "gold" label.

Patient Question: Hi I was in hospital in a critical care unit after having a fit and started shaking on one side of my body. I spaced out. I have woken up but I am confused, I don't remember what happened. I have been told that there is a shadow on my brain what could this be
Clinician-interpreted Question: What caused the patient's confusion and short-term memory loss?
System Response #1: What are possible causes of a brain shadow after seizure activity? **Quality:** *Ok* | **Error:** *Generic* **Reason:** Generated question is general.
System Response #2: What is the likely cause of the new cerebral lesion on imaging? **Quality:** *Bad* | **Error:** *Hallucination* **Reason:** Patient question does not mention a lesion.

Table 5: Complete example of patient-to-clinician question reformulation and human evaluation of model outputs.

C.2 Automatic metric distributions

Figures 4 and 5 show distributions of ROUGE-L, BERTScore, Medcon, and AlignScore stratified by human *Quality* labels, computed against the reference clinician-authored question and the patient-authored narrative, respectively.

C.3 Pairwise metric meta-evaluation

Table 6 reports pairwise accuracy results for automatic metrics against human preferences under both comparison settings.

Additional observations. Table 6 also shows a few additional trends. In the clinician question setting, Medcon yields ties for more than half of the comparisons (72/123), although its accuracy rises to 64.7% when ties are excluded. These ties likely reflect Medcon's frequent zero scores, since the compared texts are short and often share few or no clinical entities. In the patient narrative setting, Medcon still produces many ties (27/123), although fewer than in the clinician question setting. We also observe greater asymmetry by the human-preferred system in the patient narrative setting, suggesting that some metrics are more sensitive to cases where the 70B output is preferred than to cases favoring the 8B output.

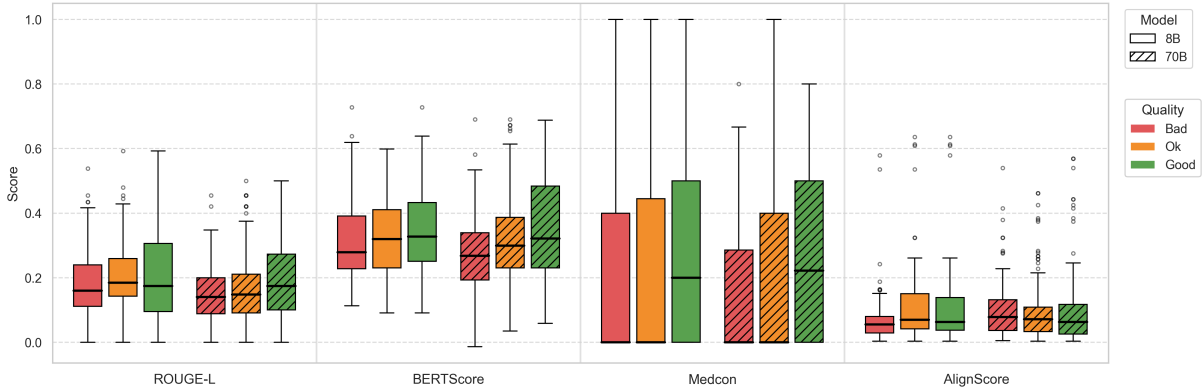


Figure 4: Distributions of automated metric scores computed between each system-generated question and the reference clinician-authored question, stratified by human *Quality* label, for outputs from Llama 8B and 70B.

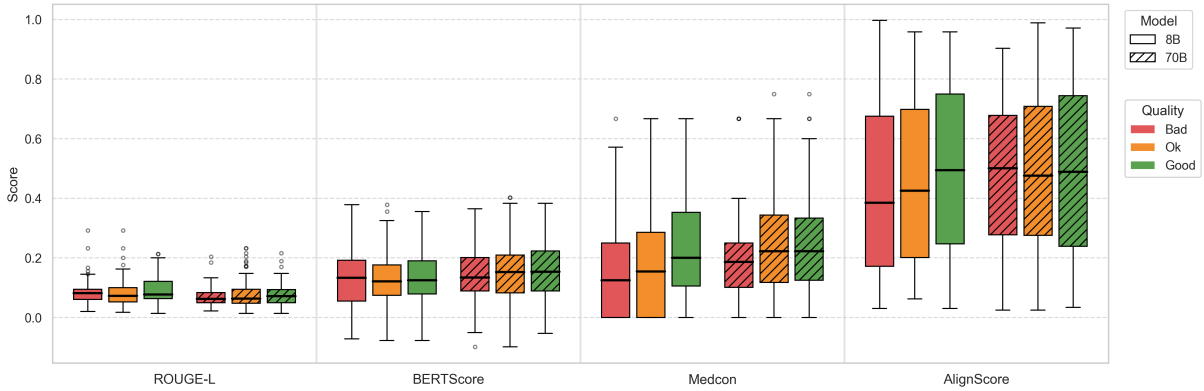


Figure 5: Distributions of automated metric scores computed between each system-generated question and the original patient-authored narrative, stratified by human *Quality* label, for outputs from Llama 8B and 70B.

Metric	Overall			PA by human winner		PA by quality pair		
	PA *N=123*	*Ties* *N=123*	*PA w/o ties* *N=123 − ties*	*8B* *N=66*	*70B* *N=57*	*good/bad* *N=28*	*good/ok* *N=58*	*ok/bad* *N=37*
Using Clinician Question								
ROUGE-L	49.6	2	50.4	51.5	47.4	60.7	44.8	48.6
BERTScore	**61.0**	0	61.0	**63.6**	**57.9**	**71.4**	**56.9**	**59.5**
Medcon	26.8	72	**64.7**	25.8	28.1	21.4	34.5	18.9
AlignScore	48.0	0	48.0	48.5	47.4	60.7	43.1	45.9
Using Patient Narrative								
ROUGE-L	48.0	4	49.6	**43.9**	52.6	**53.6**	43.1	51.4
BERTScore	**52.0**	0	52.0	39.4	**66.7**	**53.6**	**44.8**	**62.2**
Medcon	40.7	27	**52.1**	30.3	52.6	42.9	37.9	43.2
AlignScore	48.0	0	48.0	36.4	61.4	39.3	**44.8**	59.5

Table 6: Meta-evaluation of automatic metrics against human pairwise preferences on a subset of cases where the two models received different human quality ratings. PA denotes pairwise accuracy. All entries except *Ties* are percentages (%); denominators are shown below the column headers. PA treats metric ties as disagreements, whereas PA w/o ties is computed on the non-tied subset only. Columns 8B and 70B report PA conditioned on which system the human preferred, and columns good/bad, good/ok, and ok/bad report PA within each human quality-pair bucket. Bold indicates the highest overall PA within each setting.

Investigating Stigmatizing Language in Clinical Documentation with Open-Source Large Language Models

Rajashree Dahal,[1] **Pardis Hossein Pour,**[2] **Pranathi Kamisetty,**[1]
Satwik Pamulaparthy,[1] **Saeid Tizpaz-Niari,**[1] and **Natalie Parde**[1]
[1]Department of Computer Science, University of Illinois Chicago, IL, USA
[2]College of Applied Health Sciences, University of Illinois Chicago, IL, USA
{rdaha1, phoss, pkami, vpamu, saeid, parde}@uic.edu

Abstract

Clinical documentation is essential for patient care, billing, and medical research, but it is subject to entrenched bias. While manual chart reviews can identify such bias, they are labor-intensive and expert-dependent. We introduce and evaluate StigMAD, a Multi-Agent Debate framework leveraging open-source Large Language Models (LLMs) to detect stigmatizing language in clinical documentation. We investigate reasoning (multi-agent debate), self-reflection, and self-consistency within this framework. Experiments on clinical notes and patient summaries demonstrate that our framework provides significant advantages over rule-based and supervised baselines. A domain-specific LLM (MedGemma) achieved its highest performance using the StigMAD reasoning framework, while a general-purpose LLM (Llama) showed superior results with the self-consistency framework. These findings suggest that open-source LLMs, steered by structured prompting and reflective reasoning, can effectively support the auditing of stigmatizing language, marking a critical step toward more equitable clinical NLP systems.

1 Introduction

A wide variety of clinical documentation, including history and physical examination notes, daily progress notes, special consults, surgery or procedure reports, discharge summaries, and nurses' notes, can contribute to narrating patient care. Holistically, these narratives provide continuity of care across diverse medical teams and support activities such as billing and medical research (Rizvi et al., 2016). They may be supplemented by other free-text healthcare documentation such as interview transcripts, patient-provider communication logs, case reports, and survey responses, especially in research or mental health settings (Park et al., 2019). However, all of this documentation may vary in tone, structure, and completeness, and bias

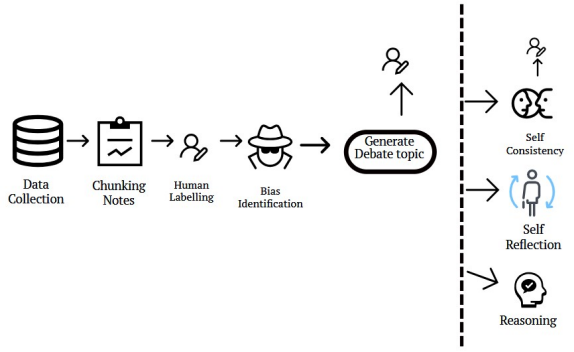


Figure 1: Methodological framework for bias detection.

can easily enter the process. For example, Black patients are more than twice as likely as white patients to have negative descriptors applied in their notes, which can influence diagnosis, treatment, and referral decisions (Sun et al., 2022). Stigmatizing language and assumptions in documentation can affect downstream AI and human decisions (Liu et al., 2023) and can produce disparities in clinical, policy, and organizational outcomes (Rozier et al., 2022). Bias can also lead to inappropriate provider decision support, lower reimbursements, and disparities in treatment recommendations, affecting clinical decision-making and contributing to health disparities in patient care (Obermeyer et al., 2019).

Bias in clinical documentation has traditionally been uncovered through manual chart review, qualitative coding, and formal interviews. All of these processes are time-consuming and expert-reliant, making them difficult to scale. To address this, researchers have started to harness artificial intelligence to uncover patterns of bias in clinical documentation (Zhang et al., 2020), for example by detecting stigmatizing language, classifying note sentiment (Valentine et al., 2024), or predicting bias-prone words. However, automatically detecting implicit forms of bias in continually evolving healthcare settings is challenging. Approaches

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 490–501
July 3-4, 2026 ©2026 Association for Computational Linguistics

often rely on fixed labeled data and miss subtle, context-dependent bias (Himmelstein et al., 2022). Frameworks relying on lexicon-based detection of these phenomena may have limited performance without further expansion and more sophisticated classification methods (Walker et al., 2025).

In this work, we introduce the STIGMAD framework, which harnesses self-consistency, self-reflection, and multi-agent debate to identify stigmatizing language using open-source LLMs, hypothesizing that the framework will enable models to reason through ambiguity and more critically analyze output. We illustrate its potential to uncover hidden stigmatizing language patterns and offer an adaptive strategy for identifying and analyzing bias in clinical documentation. Our hope is that the framework can be replicated for use with other forms of bias in the future, and we lay the groundwork for doing so in this work. Our core contributions involve (1) evaluating how effectively and confidently open-source LLMs can detect stigmatizing language in clinical documentation using the STIGMAD framework; (2) deriving how stigmatizing language observed in clinical documentation propagates into clinical summary notes; and (3) ablating individual components of the STIGMAD framework, including self-consistency, self-reasoning, and self-reflection.

2 Literature Review

2.1 Bias in Clinical Documentation

Previously, Chen et al. (2024a) systematically reviewed AI-based bias mitigation and detection strategies for electronic health records (EHRs), identifying implicit bias, selection bias, algorithmic bias, and temporal bias. FitzGerald and Hurst (2017) discussed the presence and impact of implicit biases in healthcare professionals, performing a meta-analysis of findings from 42 studies and establishing that healthcare professionals have extensive implicit biases across numerous areas, including race, ethnicity, and gender. Burton et al. (2022)'s study of 401 patient safety reports found unequal reporting by physicians' gender and race/ethnicity. They also regularly found subthemes such as "inappropriate communication," "verbal abuse," "ignoring or omission of procedure," and "physical intimidation."

Wong and Jackson (2023) discussed positive bias, negative bias, stigmatizing language, implicit bias, explicit bias, gender bias, and racial bias, and highlighted common healthcare domains in which bias may arise. Similarly, Vela et al. highlighted stigma, racial bias, explicit bias, implicit bias, and gender discrimination in their study. We focus on stigmatizing language specifically as our test case, using the subset of MIMIC-IV clinical notes annotated for "Stigmatizing Language" in Harrigian et al. (2023)'s study. Harrigian et al. (2023) identified and classified three types of stigmatizing language in medical records: *credibility and obstinacy*, *compliance*, and *descriptors*.

2.2 LLM-Mediated Bias Analyses

Several studies have examined the potential of LLMs to aid in bias detection (Qiao et al., 2024; Hayes, 2025; Eschrich and Sterman, 2024), by performing zero-shot detection and categorization of bias in emergency department and discharge notes (Apakama et al., 2024). Xavier et al. (2026) evaluated the performance accuracy of LLMs in detecting stigmatizing language across diffent temperature settings and model size. We build on these studies to explore the potential of open-source general-domain (i.e., `Llama3.3-70b-instruct` (Dubey et al., 2024)) and domain-specific (i.e., `medgemma-27b-text-it` (Sellergren et al., 2025)) LLMs to detect stigmatizing language in clinical documentation. In more general domains, counterfactual techniques (e.g., Counter-GAP (Xie et al., 2023)) have emerged as strong performers in this area by enabling precise measurement of model biases through minimally altered, attribute-specific text pairs. Approaches like FACTUAL (Li et al., 2025) further leverage counterfactual augmentation to detect and effectively mitigate bias in LLMs, particularly for stance detection tasks.

With respect to approaches leveraged in our STIGMAD framework, Madaan et al. (2023) introduced *self-refine*, which refines initial LLM output through iterative feedback and refinement. Similarly, Chen et al. (2024b) introduced *self-debugging*, where the model identifies its mistakes in the output it generates. Although several studies have revealed that self-reflection may deteriorate performance (Chen et al., 2024b; Huang et al., 2024; Valmeekam et al., 2023), Xu et al. (2024) found that LLM size can reduce bias in the self-refinement pipeline. We use models with larger parameter sizes to implement *self-reflection* in our study. Finally, *self-consistency* (prompting the LLM with the same prompt multiple times while using a higher temperature value) has produced

promising results in many domains (Ahmed and Devanbu, 2023; Wang et al., 2023).

2.3 Agentic AI for Stigmatizing Language Detection

Acharya et al. (2025) comprehensively studied the use of agentic AI across many fields, including healthcare, motivating our exploration of an agentic AI framework for stigmatizing language detection. While Du et al. (2023) explored LLMs' reasoning ability through a multi-agent debate (MAD) framework, recent studies suggest that although MAD improves reasoning, it can inadvertently reinforce societal biases. For example, Borah and Mihalcea (2024) show that implicit gender biases persist and often escalate during multi-agent LLM interactions. Oh et al. (2026) correspondingly identified bias reinforcement as a key limitation of MAD. These limitations may be attributed to a lack of context and shared reasoning patterns (Borah and Mihalcea, 2024; Oh et al., 2026). Our setup in STIGMAD addresses these limitations fundamentally. By providing each agent with a clearly defined debate paragraph, a quote highlighting the bias, and explicit information about the stigmatizing language bias and its definition, we reduce ambiguity and encourage deliberate, reflective reasoning. We hypothesize that this makes bias reinforcement less likely in our setting, since agents are tasked with detecting and debating bias explicitly.

3 Dataset

3.1 Data Sources

We downloaded the PMC-Patients summary dataset (Zhao et al., 2023) from HuggingFace and also obtained access to the MIMIC-IV Dataset (Johnson et al., 2024, 2023) of real-world clinical documents.[1] Accessing MIMIC-IV notes required credentialed access via PhysioNet (Goldberger et al., 2000) and acceptance of a Data Use Agreement. All uses are within the scope of non-commercial research, in accordance with their respective licenses and access agreements. PMC-Patients summary notes correspond to the 167,000 patient summaries from PMC-Patients including patient visit, medical history, symptoms, treatments, discharge summary, and intervention based on case reports in PubMed Central (PMC). The stigmatization subset of MIMIC-IV is a collection of 4,710 de-identified free-text clinical notes for 4,259 patients and contains 5,043 annotations.

Notes in the MIMIC-IV stigmatization subset had 1802.71±479.59 tokens on average, and summaries in PMC-Patients had 288.58±113.28 tokens. Given the length disparity, we selected seven MIMIC-IV notes and 50 PMC-Patients summaries for further annotation of stigmatizing bias. These document quantities provided us with equal numbers (n=71 each) of context length-limited chunks. The design of our reasoning framework motivated our focus on chunk-level analysis rather than full-file analysis, and the inclusion of both MIMIC-IV and PMC-Patients allowed us to validate the generalizability of our framework.

3.2 Stigmatizing Language

Stigmatizing language has been defined in numerous ways in prior work (Harrigian et al., 2023; Himmelstein et al., 2022; Barcelona et al., 2024). While some studies treat the terms "stigma," "stigmatizing language," and "bias" interchangeably (Barcelona et al., 2024), others emphasize that stigmatizing language can serve as a mechanism for transmitting bias (Himmelstein et al., 2022). Harrigian et al. (2023) positioned the detection of stigmatizing language as a bias-related task in relation to other bias detection tasks in NLP, and we adopted their definition in our work:

> ***Stigmatizing Language:*** *Any language that portrays the patient in a negative or judgmental way—by expressing doubt about their credibility, implying they are difficult, noncompliant, or deceptive, or by using language that casts their behavior, attitude, or emotional state in a negative light. This includes both overt and subtle wording that reflects blame, shame, or disrespect.*[2]

In Prompt A.1, we also included definitions of other bias types from existing literature alongside stigmatizing language to reinforce that in real-world scenarios, multiple biases may coexist.

[1] This study was reviewed by the Institutional Review Board (IRB) at our institution and determined not to involve human subjects as defined by the Department of Health & Human Services (DHHS) and/or U.S. Food and Drug Administration (FDA) regulations.

[2] The final single definition is curated using class-based definitions from three stigma types for Stigmatizing Language.

3.3 Annotation

For MIMIC-IV, we implemented the existing stigmatizing language annotations developed by Harrigian et al. (2023). For PMC-patients summary dataset, we developed annotation guidelines collaboratively with a licensed physical therapist who had extensive experience writing and analyzing clinical notes. Additional guidance was provided by experts specializing in (1) fair and responsible AI, and (2) the intersection of computer science and healthcare. The physical therapist and three computer science graduate students carried out the annotations, following the guidelines provided in §3.2 and Prompt A.1. The graduate students followed the established rules to extend coverage and ensure consistency across the dataset, while the physical therapist oversaw the work to maintain clinical accuracy. All annotators underwent a training session using real clinical notes (not included in the final dataset) to practice annotations together and ensure consistency. They separately annotated every chunk (n=71) and held follow-up meetings to discuss discrepancies and align their understanding. The experts were also involved in these meetings to ensure that clinical and methodological perspectives were incorporated.

We calculated inter-annotator agreement for PMC-Patients using Cohen's cappa (Cohen, 1960) and obtained k=0.70 which is a substantial agreement (Landis and Koch, 1977). A clinician adjudicated the final label for cases of disagreement. Stigmatizing language was found in n=28 chunks and n=19 chunks from the MIMIC-IV and PMC-Patients datasets, respectively.

4 Methodology

4.1 Baseline Framework

Given an input clinical document, our base framework *bias-identifier* (see Prompt A.1) identifies instances of stigmatizing language and uses these as debate topics. We leverage two open-source LLMs, `Llama-3.3-70B-Instruct` and `medgemma-27b-text-it`, to generate the debate topic using a multi-step process. We first ask the LLM to identify stigmatizing language, based on a relevant quote from the input. We set the LLM temperature to 0.1 to encourage more deterministic detection, maintaining logical consistency.

Since medical notes vary in size depending on note-taking context, we chunk the note into 300-token windows, with a sliding window chunk overlap of 10 tokens to preserve coherence across segments.[3] Each chunked paragraph is fed to the LLM to extract the bias (stigmatizing language) itself and quotes. We conducted this process in a zero-shot manner, experimenting with multiple prompting strategies and styles to identify those that performed best.

4.2 Self-Reflection Framework

Our self-reflection framework revisits previously identified quotes and stigmatizing language from clinical documents, along with the definition of stigmatizing language. Its role is to re-evaluate whether the quote within the context of the note truly reflects the specified stigmatizing language. The model provides a boolean output regarding its decision, without providing any explanation. Since justification and reasoning are already addressed elsewhere within the StigMAD framework, the binary reflection provides complementary validation of the correctness of the extracted quote and the predicted stigmatizing language. We used few-shot prompting to implement this framework. The prompt contains few-shot examples of other bias type from existing literature to reinforce real-word scenarios (see Prompt A.2). This framework is conceptually similar to stigma detection guided prompt by (Chen et al., 2025). However, Chen et al. (2025) rely on fixed set of predefined stigmatizing keywords whereas self-reflection uses contextually extracted quotes identified dynamically by bias identifier.

4.3 Self-Consistency Framework

Our self-consistency framework is analogous to *bias-identifier*, but we run this model eight times while increasing the temperature to 0.5 to promote response diversity. The added variability makes it easier to identify consistent outputs through voting across runs. We consider only those outputs for stigmatizing language that occur at least three times out of eight runs for a given input chunk. We keep this threshold low since the *bias-identifier* also extracts quotes, and the exact quote boundary may vary across different runs. The prompt used to implement the self-consistency framework is the same as that of *bias-identifier*.

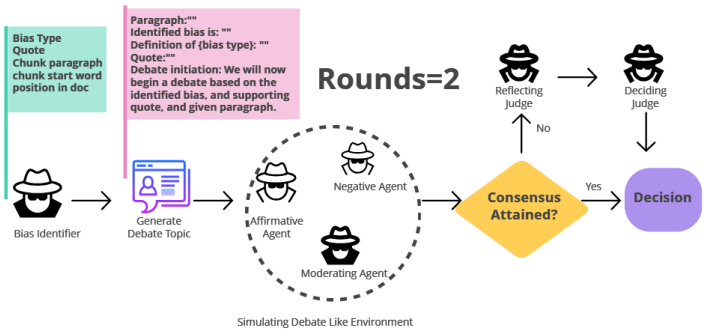


Figure 2: Proposed STIGMAD framework.

4.4 Reasoning Framework

The quotes for the stigmatizing language generated by the *bias-identifier* are converted into a debate topic as shown in Figure 2. The debate topic includes the input clinical document chunk, identified stigmatizing language, definition of stigmatizing language, quote that supports the identified bias, and debate initiation process. Overall, STIGMAD comprises four types of agents:

Affirmative Agent. Can support or deny the debate topic. It justifies its reflection on that topic, arguing affirmatively in brief structured statements (see Prompt A.4).

Negative Agent. Challenges or critiques the affirmative agent's perspective, providing arguments or support for its position (i.e., whether the stigmatizing language exists) (see Prompt A.5).

Moderator Agent. Oversees debate rounds and analyzes responses from affirmative and negative agents, determines whether consensus on bias classification has been reached, and summarizes reasons supporting its final decision. It ensures consistency across debates, reassesses contentious cases, and validates that the consensus aligns with bias definitions. The agent moderates the debate process for two rounds. If consensus is reached at the final round, its decision is the final decision (see Prompt A.3).

Judge Agent. Triggered when no consensus is reached after two moderation rounds. The judge agent takes two roles. The *reflecting judge* (see Prompt A.6) compiles upto five final arguments from affirmative and negative agents's results of the final debate rounds conclude, by listing bias classifications and reasoning without initial evaluation. The *deciding judge* (see Prompt A.7) makes the ultimate decision regarding bias presence based on its initial reflection. This agent is fed with debate topic information, and uses compiled pointers as context to produce a single JSON decision in the form `{"Presence":"True/False", "Reason":""}` which constrains output to one binary label and one justification.

4.5 External Baselines

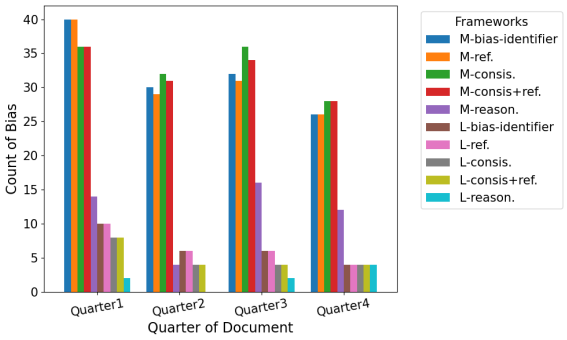
For the stigma-annotated MIMIC-IV dataset, we also compared to the rule-based baseline established by Himmelstein et al. (2022) and the supervised stigmatizing language detection model created by Harrigian et al. (2023) as an external point of comparison on the same data [4].

5 Evaluation

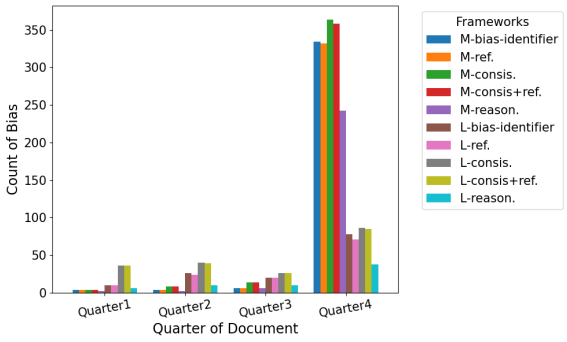
We evaluated our frameworks across PMC-Patients and MIMIC-IV. Experiments were run using four NVIDIA A100 80GB PCIe GPUs and runtime for the full suite of experiments was approximately four days. Models were loaded in native precision float (bfloat16) across four A100 GPUs. Decoding used the temperature 0.1 for bias-identifier and self-reflection frameworks, and temperature 0.5 for self-consistency and reasoning. For bias-identifier and self-consistency framework, max_length was set to 4096. Similarly, max_new_tokens was set to 1 and 250 for self-reflection and reasoning framework respectively. In the following subsections, we discuss stigmatizing language across datasets and documents (§5.1), the framework-level evaluation (§5.2), and comparison to external baselines (§5.3). For brevity, we

[3]We observed hallucination and context drift when longer passages were provided. The reasoning framework also benefited from shorter, focused spans to directly interrogate.

[4]Code is available at https://github.com/RajashreeDahal4/Stigmatizating_Language_identification.



a. PMC-Patients Summary Dataset.



b. MIMIC-IV Dataset.

Figure 3: Positional stigmatizing language distribution across datasets. M=*MedGemma*, L=*Llama*, ref.=bias-identifier+reflection framework, consis.=consistency framework, reason.=bias-identifier+reasoning framework, consis+ref.=consistency followed by reflection framework.

refer to `Llama-3.3-70B-Instruct` as *Llama* and `medgemma-27b-text-it` as *MedGemma* throughout §5.1–§5.3.

5.1 Stigmatizing Language across Datasets and Documents

We found that *bias-identifier* extracted 13 (Llama) and 64 (MedGemma) debate topics from PMC-Patients. It extracted 68 (Llama) and 178 (MedGemma) debate topics from MIMIC-IV. When dividing each note into equal-length quartiles, stigmatizing language was observed across all quartiles for PMC-Patients (Figure 3), with a slight elevation in the first quartile, whereas it was heavily concentrated in the final quartile for MIMIC-IV. MedGemma extracted more stigmatizing language across all quartiles for PMC-Patients, and in the last quartile for MIMIC-IV notes. The more frequent presence of stigmatizing language near the end of MIMIC-IV notes suggests that it may feature more prominently in evaluative language, which is usually found towards the end of healthcare provider notes. Except for reasoning, stigmatizing language concentrations were consistent regardless of which framework (see §4) was implemented using a particular LLM.

5.2 Framework Performance

We compared stigmatizing language identified by all proposed frameworks (*bias-identifier*, *self-consistency*, *self-reflection*, and *reasoning*) across both datasets using both models. We observed that the *self-reflection* and *reasoning* frameworks penalized documents or remained consistent with lower stigmatizing language counts in most cases (see Figure 3). *Self-consistency-reflection* did not show a similar trend. This may due to the voting strategy involved in making the final prediction. The main effect of self-reflection and reasoning was to prune spurious predictions when applied on top of *bias-identifier*, reducing ambiguity in detected stigmatizing language locations. Similarly, the number of debate topics identified by *bias-identifier* and *self-consistency* were comparable, perhaps due to the LLM's confidence at generating similar debate topics across different runs. Adding self-reflection to *bias-identifier* reduced this disparity by a small margin, highlighting stigmatizing language more across both datasets.

MedGemma and Llama exhibited different trends across frameworks (see Table 1). MedGemma's reasoning framework outperformed the baseline across both datasets, whereas Llama demonstrated its lowest performance with the reasoning framework. This was interesting given that MedGemma is less than 2.5 times the size of Llama, and suggests that MedGemma's pretraining data (including a diverse set of medical text) may enable it to leverage the reasoning framework to a much fuller extent in the structured clinical setting than a larger model pretrained on general-purpose data. Also, Llama's reasoning drop might be from sparse initial extractions (13 debate topics vs MedGemma's 64, where overpruning by debate agents devastates recall). The self-consistency framework in turn demonstrated superior performance for Llama. Finally, we observed that while reasoning and consistency-based prompts sometimes increased performance, the gains were not consistent across datasets (see, e.g., Llama on PMC-Patients).

From a qualitative perspective, some examples of false positive quotes included "he did not comply with during his admission," "homeless male with long history of etoh and heroin abuse," and

Table 1: Performance across models*.

Model	MIMIC(m)			PMC(p)		
	Framework	F_1	A	Framework	F_1	A
MedGemma	m-bias-identifier	0.36	0.34	p-bias-identifier	0.54	0.66
MedGemma	m-refl.	0.36	0.35	p-refl.	0.55	0.68
MedGemma	m-reason.	0.39	0.42	p-reason.	0.59	0.80
MedGemma	m-consis.	0.37	0.28	p-consis.	0.49	0.62
MedGemma	m-consis+refl.	0.37	0.28	p-consis+refl.	0.51	0.65
Llama	m-bias-identifier	0.34	0.73	p-bias-identifier	0.43	0.82
Llama	m-refl.	0.34	0.73	p-refl.	0.43	0.82
Llama	m-reason.	0.34	0.73	p-reason.	0.20	0.77
Llama	m-consis.	0.51	0.73	p-consis.	0.43	0.82
Llama	m-consis+refl.	0.51	0.73	p-consis+refl.	0.43	0.82

* F_1= F_1 score and A = accuracy. A total of 71 chunks were human-labeled, each treated as either stigmatizing or not.

"pt's paranoia seemed to be based in reality as the other resident in past had 'picked on' a gay resident." Some of these flagged examples are qualitatively ambiguous and may require closer examination in future work.

5.3 Comparison to External Baselines

We aligned Harrigian et al. (2023)'s fine-grained work on the MIMIC-IV stigmatizing language subset by mapping their "Exclude," "Neutral," and "Positive" model predictions to *non-stigmatizing* and the rest ("Disbelief," "Difficult," "Negative") to *stigmatizing*. We also fine-tuned the `emilyalsentzer/Bio_ClinicalBERT`[5] (Alsentzer et al., 2019) model on the binarized MIMIC-IV stigmatizing language data (*baseline-stigma*) using a 0.7/0.2/0.1 train-test-validation split. The 7 MIMIC-IV evaluation notes were part of test split. All reasoning frameworks outperformed these external alternatives (Figure 4) by a wide margin across models and datasets, with the exception of MedGemma with MIMIC-IV (`M-m`). This exception might be due to the poor *bias-identifier* score upon which the reasoning framework is dependent. The *bias-identifier* framework outperforms the alternatives in general though, justifying the additional computational cost of LLM-based approaches for improved performance at identifying stigmatizing language. While *baseline-stigma* slightly outperformed *L-bias-identifier*, we observed that *L-consistency* and *L-consistency + reflection* outperformed by a wider margin (13.5% improvement in precision). The *baseline-stigma* method achieved a perfect recall (1.0) for stigmatizing language but low precision (0.37).

[5]We implemented Bio_ClinicalBERT because the original work by Harrigian et al. (2023) implemented this model in their work.

6 Discussion and Conclusion

Through the study of self-reflection, self-consistency, and self-reasoning frameworks, our work sheds light on the use of LLMs to detect stigmatizing language across two popular clinical datasets. We found that *bias-identifier* tended to over-predict stigmatizing language with low precision. Self-reflection improved (or remained constant) over this outcome, and self-consistency introduced stability across multiple runs. MedGemma generally had higher performance using the STIGMAD framework across both datasets, although Llama achieved stronger performance using the self-consistency framework specifically. Our analysis shows that self-reflection and self-consistency can provide comparable results to more complex self-reasoning frameworks. Self-reasoning outperforms alternative frameworks when used with MedGemma, although not when used with the general-domain Llama. This might be because Llama, as a general-purpose model lacks clinical domain grounding to correctly adjudicate borderline cases during debate.

We focused our investigation on building comprehensive understanding of the capacity of STIGMAD to recognize stigmatizing language in clinical settings. The self-consistency framework itself validates performance stability, through its inherent aggregation of multiple sampled outputs (repeated runs) of *bias-identifier*. While STIGMAD incurs higher computational cost compared to simpler approaches (e.g., *baseline-stigma*), our study reveals what appears to be a clear performance tradeoff. For those wishing to study stigmatizing language in their own clinical environments as a downstream

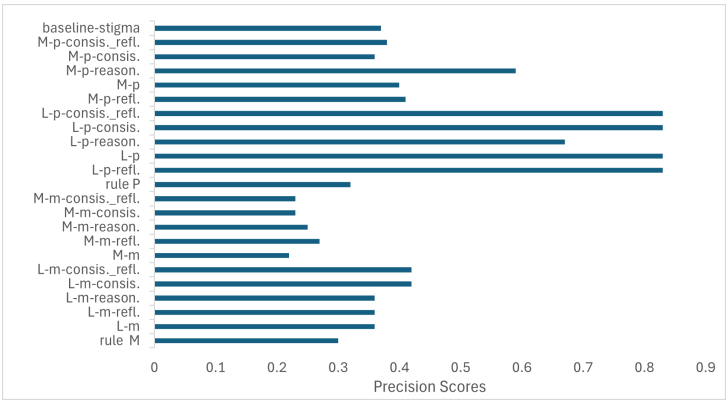


Figure 4: We compare precision scores of the rule-based method (*rule*) proposed by Himmelstein et al. (2022) and the baseline stigma detection method (*baseline-stigma*) from Harrigian et al. (2023) on the MIMIC-IV dataset, alongside our proposed frameworks evaluated on both MIMIC-IV and PMC using two LLMs. Here, `reason.` denotes reasoning, `refl.` denotes self-reflection, `consis.` denotes consistency, `p` refers to PMC, `m` refers to MIMIC-IV, `L` denotes Llama, and `M` denotes MedGemma.

application, we recommend weighing the relative performance advantages of STIGMAD against the available computational infrastructure and/or associated costs to build such an infrastructure.

Applying reasoning frameworks to stigmatizing language detection in clinical documents is an emerging topic; thus, a large part of the innovation for the work reported here lies in the task formulation and the discovery of domain-specific debate protocols. The resulting performance advantages compared to internal and external baselines highlights the strength of the implementation. Overall, our results confirm that open-source LLMs, steered by structured prompting and agent roles through STIGMAD, are well-positioned as a tool to detect stigmatizing language in clinical text. Our hope is that this work acts as a stepping stone to the development of responsible clinical NLP systems for stigmatizing language detection, as well as for other forms of bias detection and auditing.

Limitations

We aim to scale this research to larger, more heterogeneous clinical datasets to evaluate performance across more diverse clinical settings. We will also experiment with smaller, resource-light models for self-consistency and self-reflection to allow for faster and more scalable clinical deployment.

Our research was carried out on 57 patient files from MIMIC-IV (n=7) and PMC-Patients (n=50), which may not be representative of the overall datasets. We also acknowledge that annotating stigmatizing language requires interpreting tone and writer intent from text alone, which naturally opens the door to ambiguity. To mitigate the potential for this, annotators were trained to closely follow stigmatizing language definitions grounded in contemporary literature. Moreover, our annotation process was guided by experts in both the clinical and computational domains.

While individual outputs from LLMs are inherently non-deterministic, our framework improves internal consistency by structuring the generative process, supporting reproducible evaluation pipelines. Temperature values were set according to the conventions of their respective frameworks with low temperature for deterministic single-pass extraction and moderate temperature for diversity-dependent voting. Here, a formal sensitivity analysis across temperature remains an important direction for future work. The future work should also include sensitivity analysis over the self-consistency voting threshold and number of rounds for debate, which were set based on computational constraint and task-specific reasoning rather than formal hyperparameter search.

Finally, analyzing the influence of de-identification on stigmatizing language annotation and automated identification is beyond the scope of our research; however, it presents an intriguing avenue for future work. Likewise, broadening the analysis to include HIPAA-compliant closed-source models (e.g., ChatGPT for Healthcare) was infeasible for the current study due to cost and access constraints, but would present a valuable next step for external validation.

Ethical Considerations

This study was reviewed by the Institutional Review Board at the University of Illinois Chicago and determined not to be human subjects research according to Department of Health & Human Services (DHHS) or US Food and Drug Administration (FDA) regulations. While no direct patient interaction was involved, this research touches on ethically sensitive areas, including bias in healthcare. Our research includes a publicly available, de-identified clinical dataset (MIMIC-IV notes) and the PMC-Patients Summary dataset. While our goal is to identify bias in clinical notes, our findings are intended for academic analysis and model improvement, not direct clinical use. We have used open-source LLMs in our local machine to maintain privacy. However, we acknowledge the limitations of such models in reliably detecting nuanced or systemic bias, and emphasize that automated outputs should not replace human judgment in clinical settings.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. N. Parde was partially supported during this work by the National Science Foundation under Grant No. 2125411 and the National Institutes of Health under Grants R41NR020667, 1R61DA057629-01A1, and 1R01AG091762-01. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. 2025. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13:18912–18936.

Toufique Ahmed and Premkumar Devanbu. 2023. Better patching using llm prompting, via self-consistency. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1742–1746.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Donald U. Apakama, Kim-Anh-Nhi Nguyen, Daphnee Hyppolite, Shelly Soffer, Aya Mudrik, Emilia Ling, Akini Moses, Ivanka Temnycky, Allison Glasser, Rebecca Anderson, Prathamesh Parchure, Evajoyce Woullard, Masoud Edalati, Lili Chan, Clair Kronk, Robert Freeman, Arash Kia, Prem Timsina, Matthew A. Levin, and 8 others. 2024. Identifying and characterizing bias at scale in clinical notes using large language models. *medRxiv*.

Veronica Barcelona, Danielle Scharp, Betina R. Idnay, Hans Moen, Kenrick Cato, and Maxim Topaz. 2024. Identifying stigmatizing language in clinical documentation: A scoping review of emerging literature. *PLOS ONE*, 19(6):e0303653.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Élan Burton, Brenda Flores, Barbara Jerome, Michael Baiocchi, Yan Min, Yvonne A Maldonado, and Magali Fassiotto. 2022. Assessment of bias in patient safety reporting systems categorized by physician gender, race and ethnicity, and faculty rank: a qualitative study. *JAMA Network Open*, 5(5):e2213234–e2213234.

Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. 2024a. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5):1172–1183.

Hongbo Chen, Myrtede Alfred, and Eldan Cohen. 2025. Efficient detection of stigmatizing language in electronic health records via in-context learning: comparative analysis and validation study. *JMIR Medical Informatics*, 13(1):e68955.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024b. Teaching large language models to self-debug. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and

510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

James Eschrich and Sarah Sterman. 2024. A framework for discussing llms as tools for qualitative analysis. *arXiv preprint arXiv:2407.11198*.

Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):19.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Euee Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. 2023. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329.

Adam S Hayes. 2025. "conversing" with qualitative data: Enhancing qualitative research through large language models (llms). *International Journal of Qualitative Methods*, 24:16094069251322346.

Gracie Himmelstein, David Bates, and Li Zhou. 2022. Examination of stigmatizing language in the electronic health record. *JAMA Network Open*, 5(1):e2144967–e2144967.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). RRID:SCR_007345.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Adam Gayles, Ahmad Shammout, Steven Horng, Tom J. Pollard, Sixiang Hao, Benjamin Moody, Benjamin Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2025. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7075–7092, Albuquerque, New Mexico. Association for Computational Linguistics.

Yizhi Liu, Weiguang Wang, Guodong Gao, and Ritu Agarwal. 2023. The impact of stigmatizing language in ehr notes on ai performance and fairness. In *Proceedings of the International Conference on Information Systems (ICIS)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. 2026. From belief entrenchment to robust reasoning in llm agents. *Preprint*, arXiv:2503.16814.

J. Park, D. Kotzias, P. Kuo, R.L. LoganIV, K. Merced, S. Singh, M. Tanana, E. KarraTaniskidou, J.E. Lafata, D.C. Atkins, M. Tai-Seale, Z.E. Imel, and P. Smyth. 2019. Detecting conversation topics in primary care office visits from transcripts of patient–provider interactions. *Journal of the American Medical Informatics Association*, 26(12):1493–1504.

Shan Qiao, Xingyu Fang, Camryn Garrett, Ran Zhang, Xiaoming Li, and Yuhao Kang. 2024. Generative ai for qualitative analysis in a maternal health study: Coding in-depth interviews using large language models (llms). *medRxiv*.

Rubina F. Rizvi, Kathleen A. Harder, Gretchen M. Hultman, Terrence J. Adam, Michael Kim, Serguei V.S. Pakhomov, and Genevieve B. Melton. 2016. A comparative observational study of inpatient clinical note-entry and reading/retrieval styles adopted by physicians. *International Journal of Medical Informatics*, 90:1–11.

Michael D. Rozier, Kavita K. Patel, and Dori A. Cross. 2022. Electronic health records as biased tools or tools against bias: A conceptual model. *The Milbank Quarterly*, 100(1):134–150.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall,

and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record. *Health Affairs*, 41(2):203–211.

Alissa A. Valentine, Lauren A. Lepow, Alexander W. Charney, and Isotta Landi. 2024. The point of view of a sentiment: Towards clinician bias detection in psychiatric notes. *CoRR*, abs/2405.20582.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43(1):477–501.

Andrew Walker and 1 others. 2025. Care-sd: classifier-based analysis for recognizing provider stigmatizing and doubt marker labels in electronic health records: model development and validation. *Journal of the American Medical Informatics Association*, 32(2):365–374.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Christopher J Wong and Sara L Jackson. 2023. *The Patient-Centered Approach to Medical Note-Writing*.

Teenu Xavier, Jane M Carrington, and Joshua Lambert W. 2026. Detecting stigmatizing language with large language models: mind the settings. *JAMIA Open*, 9(2):ooag037.

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-GAP: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3761–3773, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10 1:909.

A Prompt Templates

Prompt A.1: Extracting Bias Type and Quote for Debate Topic

You are tasked with identifying if any of the following types of bias are present in the given medical note. If so, name the bias type and provide the quote that highlights the bias type in the given medical note. The definitions for bias type are:

{other bias type definitions. . . }

Stigmatizing Language – Stigmatizing language is any language in clinical documentation that portrays the patient in a negative or judgmental way—by expressing doubt about their credibility, implying they are difficult, noncompliant, or deceptive, or by using language that casts their behavior, attitude, or emotional state in a negative light. This includes both overt and subtle wording that reflects blame, shame, or disrespect. Example: "claims to see a therapist".

Note: You are supposed to answer the bias and the corresponding quote which highlights the bias in the form: `[{"bias_type":"", "quote":""},{"bias_type":"", "quote":""}]`. Please write as many bias type as you can.

Note: Your output should explicitly be in the above mentioned format and do not give any other justification. If no bias is found, just answer `[{"bias_type":"", "quote":""]}`. Do not repeat the same bias type with the same quote. Make sure your answer starts with only one "[" and ends with only one "]". You can add as many unique bias types and quote pairs within the brackets in the given format.

Medical Note:

`##medical_note##`

Answer:

Prompt A.2: Prompt used for self-reflection framework

You are a medical language model trained to identify bias in clinical notes. You previously predicted bias type in the input paragraph and the corresponding quote that highlights the bias. Now reflect on that decision. Based on the definition for bias type and input paragraph, was your prediction for bias type correct? Use the information to determine whether the quote, as it appears in the paragraph, matches the bias_type. You are supposed to answer a single digit 1 if there exists bias mentioned above, else answer single digit 0. Your answer should explicitly be a single digit 1 or 0. Do not explain or justify your answer.

###Note####: Ensure that you do not give any justification and reasoning for your answer.

###Example1:

****quote****: "The patient is non-compliant with their medication regimen."

****Paragraph****: "Despite multiple counseling sessions, the patient is non-compliant with their medication regimen and shows little interest in improving their condition."

****bias type****: Explicit Bias

**** The definition of Explicit Bias is ****: Explicit forms of bias include preferences, beliefs, and attitudes of which people are generally consciously aware, endorsed, and can be identified and communicated (e.g., labeling a patient as

'non-compliant' or 'irresponsible').)
Answer: 1
Example2:
****quote****: "feeling tired during the day."
****Paragraph****: "The patient reports feeling tired during the day but denies any issues with sleep or recent lifestyle changes. Vitals are within normal range."
****bias type****: Explicit Bias
**** The definition of Explicit Bias is ****: Overtly negative or judgmental language used in a medical note (e.g., labeling a patient as 'non-compliant' or 'irresponsible')
Answer: 0
Question:
****quote****: ##quote##
****Paragraph****: ##paragraph##
****bias type****: Stigmatizing Language
**** The definition of Stigmatizing Language is ****:
##definition##.
Answer:

Prompt A.3: Prompt for Moderator

You are a moderator overseeing a multi-agent debate involving medical bias detection. Two expert agents will analyze the same medical note and debate whether the given bias type is present, and if so, which type and why. After each round, you must assess both perspectives and decide if a consensus is reached.

Prompt A.4: Prompt for Affirmative Agent

The debate topic is ##debate_topic##
It was previously predicted that the given bias type in the input paragraph and the corresponding quote highlights the bias. Now reflect on that decision with your reasoning. Based on the definition for bias type and input paragraph, was the prediction for bias type correct? Please provide your single reasoning in less than 3 sentences. Your reasoning should be focused on the mentioned bias type. Do not introduce new type of bias. You are not supposed to ask any questions.

Prompt A.5: Prompt for Negative Agent

##aff_ans##
You are supposed to disagree with my answer. Please provide your single reasoning in less than 3 sentences. Your reasoning should be focused on the mentioned bias type. Do not introduce new type of bias. You are not supposed to ask any questions.

Prompt A.6: Prompt for Judge Reflection

Affirmative side's answer: ##aff_ans##
Negative side's answer: ##neg_ans##
Write down key pointers addressed by different experts that could be crucial to reflect on before making judgement. You are not supposed to introduce your personal pointers. Try to limit the pointers to as low as 5 sentences.

Prompt A.7: Prompt for Judge Decision

Now, based on the debate topic:
##debate_topic## Provide only one reasoning and the final decision on whether the given bias exists, and explanation. Respond strictly in JSON format: `{"Presence":"True/False","Reason": ""}`. You should provide the result in a single JSON form. Do not give more than one json answer. Your response should be explicitly in the given json format where json key and values are double quoted. You are not allowed to give any more clarification apart from the single json form.

Learning to Combine AI Annotations for Improved Biomedical Relevance Labeling

Won G Kim, Lana Yeganova, Shubo Tian,
Donald C Comeau, W John Wilbur, Zhiyong Lu
Division of Intramural Research (DIR), NLM, NIH, Bethesda, MD USA 20894
{wonkim,lana.yeganova,shubo.tian,donald.comeau,john.wilbur,zhiyong.lu}@nih.gov

Abstract

Accurate labeling of relevance between biomedical abstracts is essential for improving information retrieval, semantic similarity modeling, training of ranking systems and other Natural Language Processing tasks. However, manual annotations are time-consuming, labor intensive and costly. Studies show that large language models (LLMs) can facilitate automated annotation, but their performance still falls short of human expert-level accuracy, especially in domain-specific tasks.

It has been shown that combining annotations from multiple non-expert annotators can achieve performance comparable to, or even exceeding, that of trained experts. Based on this evidence, we treat AI-generated annotations as contributions from non-expert annotators and combine them using Learning to Rank framework. Our results show significant improvement in overall annotation quality. The proposed method looks promising to reduce reliance on human annotation while maintaining reliable performance for large-scale biomedical applications.

1 Introduction

Recent studies have examined the effectiveness of large language models (LLMs) for annotation and relevance labeling in both general NLP and biomedical domains (Lu et al., 2024; Jin et al., 2023b). In computational linguistics, LLM annotations have been shown to effectively complement human annotation workflows (Gligoric et al., 2025). In biomedical contexts, LLMs have demonstrated strong performance in structured tasks such as clinical information extraction, but still require careful prompt design and iterative refinement to achieve reliable results (Hein et al., 2025). However, multiple studies emphasize that LLM-generated annotations remain imperfect, exhibiting biases and inconsistencies that necessitate human oversight or validation (Xia et al., 2025; Tan et al., 2024). In complex scientific tasks, including systematic review and evidence synthesis, human experts continue to outperform LLMs, highlighting limitations in domain-specific reasoning and reliability (Sollini et al., 2025; Wang et al., 2024). Overall, current evidence suggests that LLMs can approximate non-expert human annotators but fall short of expert-level performance, motivating their use as efficient yet noisy annotators rather than replacements for human expertise.

Wilbur (1996) showed the pooled results of the untrained judges were almost as good as the pooled results for the trained judges and better than any single trained judge. Motivated by this finding, we treat annotations generated by LLMs as noisy and imperfect signals. Analogous to untrained annotators, in this work, we aggregate outputs from multiple LLM judgments using diverse prompts to improve annotation quality.

Wilbur and Kim (2006) studied probabilistic methods for weighting annotators based on global statistical inference from pooled judgments, showing that aggregating multiple annotations improves the prediction of held-out labels and leads to more reliable gold standards. However, their framework treats human annotations as absolute labels rather than subjective or relative assessments. For example, when two annotators assign different relevance scores to the same document pair, this is considered a disagreement, without accounting for potential agreement in ranking.

If two annotators produce perfectly aligned rankings over a set of document pairs, they may be considered fully equivalent from a rank-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 502–507
July 3-4, 2026 ©2026 Association for Computational Linguistics

ing perspective. In this study, we adopt the latter view by emphasizing agreement in ranking. This perspective is particularly aligned with traditional and modern information retrieval methods, such as BM25 (Robertson and Zaragoza, 2009) and embedding-based models such as MedCPT (Jin et al., 2023a). To incorporate such heterogeneous relevance signals, we move beyond absolute agreement and instead employ a pairwise learning-to-rank model based on gradient boosted decision trees using XGBoost (Chen and Guestrin, 2016; Burges et al., 2005), enabling effective training over diverse scoring methods.

In evaluating how well a method aligns with a gold standard, a variety of metrics have been proposed. In traditional information retrieval, Mean Average Precision (MAP) (Buckley and Voorhees, 2000) and Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) are among the most widely used measures. For assessing inter-annotator agreement, metrics such as Cohen's kappa and Spearman's rank correlation are commonly applied. MAP and NDCG place strong emphasis on top-ranked documents, making them inherently biased toward agreement at higher ranks. Cohen's kappa avoids this top-rank bias, but measures absolute agreement and does not account for consistency in ranking. In contrast, Spearman's rank correlation captures agreement in relative ordering and is not affected by bias of the top-rank or differences in absolute annotation scales, making it more suitable for evaluating ranking consistency.

2 Data Sets

RELISH is one of the largest expert-annotated dataset for article similarity, containing approximately 196,000 annotated pairs of PubMed abstracts derived from about 3,200 query articles. For each query article, the retrieved documents are assigned relevance scores on a three-point scale (1–3) that correspond to irrelevant, partially relevant and relevant.(Brown et al., 2020).

NCBI Dataset (Wilbur, 1996) consists of 5,000 PubMed abstract pairs derived from 100 query abstracts, each annotated on a 1–4 relevance scale by 13 judges (seven trained judges and six untrained judges).

3 Method

We consider seven annotation methods, including three text similarity approaches (BM25, MedCPT, and LitSense2 (Yeganova et al., 2025)) and four AI-based variants derived from different prompting strategies (Prompt 1–4). We denote these methods as F1–F7, corresponding to Prompt 1–4, BM25, MedCPT, and LitSense2, respectively. For AI-based annotation, we use GPT-5.1 and the actual prompts used for generating AI annotations are provided in the appendix. Due to the cost of annotation, we sample 272 queries from the test set defined by (Zhang et al., 2022), a subset of RELISH. This subset provides a practical balance between annotation cost and maintaining a sufficiently large sample for reliable evaluation. Hereafter, "the dataset" denotes the 272 queries together with their retrieved, annotated abstracts. As mentioned in the Introduction, given a query, our goal is to learn a model that aligns the ranking it produces with human judgments as closely as possible. To achieve this, we adopt a pairwise learning-to-rank (LTR) approach (Chen and Guestrin, 2016; Burges et al., 2005) rather than an NDCG-based LTR method commonly used in information retrieval.

4 Results

We randomly split the dataset into three folds, using two folds for training and the remaining one for testing. To show the qualities of individual AI rankers in the training and test sets, we compute the Spearman correlation for each individual AI ranker against the corresponding human judgments, as shown in Table 1 and Table 2.

Also, to assess how well individual untrained annotators align with trained experts, we use the NCBI dataset. Let T denote the set of seven trained annotators and U the set of six untrained annotators. For each trained annotator $t \in T$ and untrained annotator $u \in U$, let $\rho_{t,u}$ denote the Spearman correlation between their annotations. First, we average correlations across trained annotators for each

Feature	Spearman
F1	0.326
F2	0.320
F3	0.354
F4	0.397
F5	0.280
F6	0.314
F7	0.362
Average	0.336
Std. Dev.	0.035

Table 1: Spearman correlation scores for individual features for the training set.

Feature	Spearman
F1	0.337
F2	0.351
F3	0.376
F4	0.407
F5	0.277
F6	0.327
F7	0.372
Average	0.350
Std. Dev.	0.042

Table 2: Spearman correlation scores for individual features for the test set.

untrained annotator:

$$\rho_u = \frac{1}{|T|} \sum_{t \in T} \rho_{t,u} \tag{1}$$

We then average these values across all untrained annotators to obtain the overall agreement between untrained and trained annotators, denoted $\bar{\rho}_U$:

$$\bar{\rho}_U = \frac{1}{|U|} \sum_{u \in U} \rho_u \tag{2}$$

The results for ρ_u for $u \in U$ and $\bar{\rho}_U$ are shown in Table 3.

From Tables 1 - Table 3, untrained human annotations may outperform both AI and traditional retrieval methods, while exhibiting similar standard deviations across the three sets. This comparison between NCBI data and RELISH data has two limitations. First, judgments on the RELISH data are on a scale from 1-3, for NCBI dataset they are on a scale from 1-4. Second, instructions to the judges on the RELISH task defined relevance as "topically relevant to the seed article; within the same specific sub-field of research, i.e. an article that would be interesting to read further or could have been cited within the original work" (Brown et al., 2020). Instructions for the NCBI data task were to "judge a document relevant to a query if and only if you would wish to read the full document if you were given the task of writing the query document."(Wilbur, 1996)

Next, we apply Learning-to-Rank (LTR) with a leave-one-out (LOO) strategy to the training set to select an optimal combination of features. We evaluated all possible combinations of the seven features (i.e., $2^7 - 1 = 127$)

Judges	ρ_u
u_1	0.4371
u_2	0.3764
u_3	0.3901
u_4	0.3972
u_5	0.4072
u_6	0.3233
$\bar{\rho}_U$	0.389
Std. Dev.	0.038

Table 3: Spearman correlations between trained and untrained annotators.

using the Spearman correlation measure. Table 4 shows the top 7 feature sets ranked by Spearman correlation, with the subset [F1, F2, F3, F4, F7] achieving the highest performance and selected as the optimal combination. With the optimal feature combination [F1, F2, F3, F4, F7], we retrain the model on the training set to obtain the final optimal model $M_{optimal}$. At this stage, we emphasize that the test set remains completely unseen during the model selection and training process. We then apply $M_{optimal}$ to the held-out test set to obtain predicted scores, achieving a Spearman correlation of 0.460, which markedly outperforms all individual features on the test set, with differences that are statistically significant (see Table 5).

Similarly to the evaluation of untrained human annotators against trained experts, we assess how well trained annotators agree with each other, using the same overall agreement measure defined for the untrained annotators. Let T denote the set of seven trained annotators. For any pair of distinct annotators $t, t' \in T$ with $t \neq t'$, let $\rho_{t,t'}$ denote the Spearman correlation between their annotations. For

Feature Set	Spearman
[F1,F2,F3,F4,F7]	0.4249
[F1,F2,F3,F4,F6,F7]	0.4248
[F1,F3,F4,F7]	0.4231
[F1,F2,F3,F4,F5,F7]	0.4217
[F1,F3,F4,F5,F6,F7]	0.4207
[F3,F4,F7]	0.4200
[F1,F2,F3,F4]	0.4197
[F2,F3,F4,F7]	0.4197

Table 4: Top 7 feature combinations selected from all 127 subsets using leave-one-out (LOO) Learning-to-Rank, ranked by Spearman correlation on the training set.

Feature	Spearman	p-value
F1	0.337	4.57e-08
F2	0.351	4.29e-07
F3	0.376	7.37e-07
F4	0.407	4.05e-04
F5	0.277	1.15e-12
F6	0.327	8.62e-10
F7	0.372	2.13e-09
$\mathbf{M}_{optimal}$	**0.460**	–

Table 5: Spearman correlation on the test set and Wilcoxon signed-rank test p-values comparing individual feature-based methods against the final model $M_{optimal}$.

each annotator $t \in T$, we first average the correlations with all other annotators:

$$\rho_t = \frac{1}{|T|-1} \sum_{t' \in T, t' \neq t} \rho_{t,t'}. \quad (3)$$

We then average these values across all trained annotators to obtain the overall agreement between trained annotators, denoted $\bar{\rho}_T$:

$$\bar{\rho}_T = \frac{1}{|T|} \sum_{t \in T} \rho_t. \quad (4)$$

The results for ρ_t for $t \in T$ and $\bar{\rho}_T$ are shown in Table 6. The overall Spearman correlation is $\bar{\rho}_T = 0.453$, which is lower than the Spearman correlation obtained by the model $M_{optimal}$ (0.460). This suggests that the model $M_{optimal}$ may achieve annotation quality comparable to that of human experts.

5 Conclusions

We explored seven heterogeneous annotation scores, comprising four LLM-based annotations generated using four different prompts and three scores derived from traditional retrieval methods, and investigate how to best combine them. Based on comparisons with human annotators, we find that each individual feature score used in our study falls short of untrained human performance. An LTR model with pairwise ranking optimization is then employed to select and combine the most effective feature combination, yielding an optimal model. The resulting model significantly outperforms each individual feature scores. Furthermore, the model may achieve annotation quality comparable to that of human experts.

Judges	ρ_t
t_1	0.4576
t_2	0.4374
t_3	0.4552
t_4	0.5085
t_5	0.3485
t_6	0.4850
t_7	0.4815
$\bar{\rho}_T$	0.4534
Std. Dev.	0.052

Table 6: Spearman correlations among trained annotators.

6 Limitations and Future Work.

Our evaluation uses 272 RELISH queries due to LLM annotation cost, which may limit generalization. Comparisons with NCBI annotations are also imperfect due to differences in relevance scales, document pairs and annotator instructions. In future work, we will scale evaluation using multiple open-source LLMs instead of a single model to study a larger query set. We also aim to investigate whether aggregating or pooling outputs from multiple weaker-performing LLMs can achieve performance comparable to stronger, more expensive models. We will see how well our optimal model performs on the NCBI data.

Acknowledgments

This research was supported [in part] by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings

and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2020. A large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2020:baaa039.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *ICML*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident llm annotations be used for confident conclusions? In *Proceedings of NAACL 2025*.

David Hein, Alana Christie, Michael Holcomb, and 1 others. 2025. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *npj Digital Medicine*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. In *Proceedings of SIGIR*, pages 41–48.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023a. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2023b. Retrieve, summarize, and verify: How will chatgpt affect information seeking from the medical literature? *Journal of the American Society for Information Science and Technology*.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024. Large language models in biomedicine and health: Current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9):1801–1811.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Martina Sollini, Cristiano Pini, Alexandra Lazar, Fabrizia Gelardi, Gaia Ninatti, Matteo Bauckneht, Arturo Chiti, and Margarita Kirienko. 2025. Human researchers are superior to large language models in writing a medical systematic review in a comparative multitask assessment. *Scientific Reports*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Z Wang, L Cao, B Danek, Q Jin, Z Lu, and J Sun. 2024. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv:2406.17755*.

W. John Wilbur. 1996. The knowledge in multiple human relevance judgments. *ACM Transactions on Information Systems*, 14(2):150–163.

W. John Wilbur and Won Kim. 2006. Improving gold standard annotation for biomedical information retrieval. In *Proceedings of the SIGIR Workshop on Biomedical Information Retrieval*.

Meng Xia, Shradha Maharjan, and 1 others. 2025. Syncode: Synergistic human–llm collaboration for enhanced data annotation. *Information*.

Lana Yeganova, Won Kim, Shubo Tian, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2025. Litsense 2.0: Ai-powered biomedical information retrieval with sentence and passage level knowledge discovery. *Nucleic Acids Research*, 53(W1):W361–W368. Open Access.

Li Zhang, Wei Lu, Haihua Chen, Yong Huang, and Qikai Cheng. 2022. A comparative evaluation of biomedical similar article recommendation. *Journal of Biomedical Informatics*, 131:104106.

A AI Prompts

PROMPT1 = "You are an expert on the biomedical literature. Here I give you two excerpts, labelled passage A and passage B. Tell me if the entities mentioned in passage A are also being discussed in passage B. If the entities from passage A are discussed in passage B tell me if passage B seems to be saying the same thing about them or if it is just too unclear to say. After you have analyzed the passages give a final answer as

an integer: 3 if passage B says all that A says and possibly more; 2 if only part of what A says is also said in B; 1 if passage A and passage B appear to be completely unrelated. Return only a single integer from 1 to 3. Do not include explanations."

PROMPT2 ="You are an expert in biomedical literature analysis. You will receive two text excerpts labeled Passage A and Passage B. For each passage, internally determine: Sentence 1: the primary research problem addressed by the study. Sentence 2: the primary method or approach used to address that problem. Do not output these sentences; use them only for internal reasoning. Then compare Passage A and Passage B based on these internally derived sentences and output exactly one integer according to the rules below: 3: Passage B fully covers the same problem and method as Passage A, and may include additional information. 2: Passage B overlaps with Passage A in either the problem or the method, but not fully. 1: Passage A and Passage B are unrelated in both problem and method. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."

PROMPT3="You are an expert in biomedical literature analysis. You will receive two text excerpts labeled Passage A and Passage B. For each passage, internally determine the following: Sentence 1: Summarize Passage A in one sentence describing the main problem this study is trying to solve. Sentence 2: Summarize Passage B in one sentence describing the main problem this study is trying to solve. Do not output these sentences; use them only for internal reasoning. Then compare Passage A and Passage B based on these internally derived sentences and output exactly one integer according to the rules below: 3: Sentence 2 fully covers the same problem to solve as Sentence 1 and may include additional information. 2: Sentence 2 overlaps with Sentence 1 in the problem to solve, but not fully. 1: Sentence 1 and Sentence 2 are unrelated in the problems they aim to solve. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."

PROMPT4 ="You are an expert in biomedical literature analysis. You will receive two paragraphs labeled Paragraph A and Paragraph B. For each paragraph, internally determine the main conclusions being addressed in one sentence. Do not output these sentences; use them only for internal reasoning. Then compare Paragraph A and Paragraph B based on these internally derived problems and output exactly one integer according to the rules below: 3: Paragraph B addresses the same main problem as Paragraph A and may include additional information. 2: Paragraph B partially overlaps with the main problem addressed in Paragraph A. 1: Paragraph B addresses a different problem and is not relevant to Paragraph A. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."

When Does Retrieval Beat Direct LLM Diagnosis in Rare Disease? An Empirical Study of Ontology Coverage

Mohamed Elmofty
Humboldt-Universität zu Berlin
mohamed.elmofty@hu-berlin.de

Ulf Leser
Humboldt-Universität zu Berlin
ulf.leser@hu-berlin.de

Abstract

Recent high-complexity agentic systems such as DeepRare perform strongly on rare disease diagnosis benchmarks, but it remains unclear when gains come from structured knowledge access and when they come from parametric LLM knowledge. We compare phenotype-based retrieval, LLM reranking, and unrestricted LLM diagnosis across seven benchmarks covering 10,382 cases. We find a clear performance crossover driven by retrieval coverage—the fraction of cases whose true diagnosis is within the retriever's top-50: on high-coverage datasets, ontology-based retrieval dominates; on low-coverage datasets, open-ended LLM diagnosis takes the lead. Building on this, adding an LLM reranker over retrieved candidates further improves accuracy across our patient-case benchmarks, closing most of the remaining gap to agentic systems (within 2 pp on MME and LIRICAL). We trace the crossover to two structural failure modes of ontology-based retrieval—annotation sparsity and phenotypic homogeneity—and show that aggregate scores across mixed benchmarks can hide these qualitatively different diagnostic settings. These findings motivate per-dataset evaluation and hybrid diagnostic systems that combine retrieval, reranking, and parametric LLM generation based on case characteristics.

1 Introduction

Rare diseases collectively affect over 300 million people worldwide (Wakap et al., 2020), yet patients frequently endure a "diagnostic odyssey" exceeding five years (Schieppati et al., 2008). The sheer number of rare diseases—over 10,000 with known genetic etiology (Haendel et al., 2020)—and limited clinician familiarity with any individual condition make automated diagnostic support a pressing need.

Phenotype-driven diagnosis tools address this by matching patient symptoms, encoded as Human Phenotype Ontology (HPO) terms (Köhler et al., 2021), against curated disease–phenotype annotations. PhenoDP (Wen et al., 2025), a recent deep-learning-based toolkit, combines information-content (IC)-based, phi-coefficient-based, and semantic similarity measures over the HPO Annotation (HPOA) knowledge base to rank candidate diseases, achieving state-of-the-art results among phenotype-based methods.

At the other end of the complexity spectrum, the DeepRare system (Zhao et al., 2026) represents a multi-agent architecture integrating over 40 specialized tools, more than 10 biomedical knowledge sources, web-scale retrieval, and self-reflective reasoning to achieve 57.2% average Recall@1 across seven rare-disease benchmarks. This raises a natural question: *how much of this performance comes from querying curated disease–phenotype ontologies versus relying on knowledge encoded in model parameters?*

We investigate this through two complementary experiments. First, we evaluate a simple two-stage pipeline—PhenoDP retrieval followed by LLM reranking—and compare it against both unrestricted LLM diagnosis and DeepRare's per-dataset results. On the high-coverage patient-case benchmarks, the reranking pipeline is competitive with DeepRare (70.0% vs. 70.0% on MME; 54.1% vs. 56.0% on LIRICAL), despite using only two components and no runtime knowledge search beyond PhenoDP's pre-indexed HPOA annotations. On the DDD (Deciphering Developmental Disorders) disease-profile benchmark, retrieval also scores higher than DeepRare's reported number (59.0% vs. 43.0%), but DDD consists of disease-level HPO bundles rather than patient cases and overlaps ontologically with HPOA, so we treat this more as an upper bound on how well ontology-aligned benchmarks can be matched than as a head-to-head capability comparison (§3.5).

Second, comparing the two diagnostic approaches—ontology-based retrieval and unre-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 508–518
July 3-4, 2026 ©2026 Association for Computational Linguistics

stricted LLM generation—across seven datasets reveals a *performance crossover*: which approach wins is determined by *retrieval coverage*, the fraction of patient presentations for which PhenoDP successfully places the true diagnosis in its top-50 candidates. Coverage reflects whether the retriever can successfully match a patient's HPO terms to the correct disease; it fails primarily when the target disease is sparsely characterized in the HPOA knowledge base, and secondarily when the patient's phenotype set is small or when the target disease is phenotypically indistinguishable from related conditions. On high-coverage datasets, ontology-based retrieval dominates; on low-coverage datasets, the unrestricted LLM takes the lead. This crossover is invisible in aggregate metrics.

This paper makes four contributions. First, we show that LLM reranking improves on pure ontology-based retrieval on every patient-case benchmark we evaluate; on the high-coverage benchmarks MME (70.0% vs. 70.0%) and LIRICAL (54.1% vs. 56.0%), an open-weight two-component pipeline reaches numbers comparable to DeepRare, although the small sample sizes (n=40 for MME) and unmatched evaluation conditions mean these comparisons should be read as approximate rather than head-to-head (§3.5, Limitations). Second, we observe a performance crossover between ontology-based retrieval and unrestricted LLM diagnosis that tracks retrieval coverage: in our seven benchmarks, datasets above 70% coverage favor retrieval and datasets below 60% favor unrestricted LLM generation. Because no benchmark in our study falls in the 60–70% transition zone, the boundary should be read as a hypothesis consistent with these observations rather than a located threshold. Per-dataset comparisons are supported by McNemar tests and bootstrap CIs. Third, through error analysis of non-retrievable cases, we identify annotation sparsity and phenotypic homogeneity as the two structural failure modes that determine where ontology-based retrieval breaks down. Finally, an oracle analysis shows that the two approaches succeed on largely different cases, and an empirical evaluation of a *pooled* score-based threshold router shows that a single retriever-confidence threshold pooled across datasets fails to close the oracle gap—pointing to input-modality features, or to per-dataset thresholding which we did not evaluate, as candidate routing signals.

2 Related Work

Automated diagnosis for rare diseases has been pursued along two tracks: phenotype-based retrieval tools that operate over curated ontologies, and more recently, general-purpose LLMs applied to clinical reasoning. Our work connects these two lines.

On the retrieval side, tools such as Exomiser (Smedley et al., 2015), LIRICAL (Robinson et al., 2020), and PubCaseFinder (Fujiwara et al., 2018) compute semantic similarity between a patient's HPO terms and the annotations of candidate diseases in the HPOA knowledge base (Köhler et al., 2021); PubCaseFinder additionally indexes published case reports to widen its evidence base. PhenoDP (Wen et al., 2025), which we use as the retrieval backbone in this work, continues this line by combining information-content-based, phi-coefficient-based, and learned semantic similarity. Ontology-anchored methods of this kind are bounded by the completeness of HPOA annotations, so diseases or cases that are under-annotated in HPOA or phenotypically similar to many others are systematically hard to retrieve.

Retrieve-and-rerank pipelines are a standard pattern in open-domain QA and information retrieval, where a fast recall-oriented retriever is followed by a stronger but more expensive reranker (Karpukhin et al., 2020; Nogueira and Cho, 2019). We apply this pattern to ontology-based rare disease diagnosis.

On the LLM side, general-purpose models have been evaluated for clinical reasoning both closed-book (Singhal et al., 2023; Nori et al., 2023) and in rare-disease-specific settings (Chen et al., 2024; Shyr et al., 2024). Domain-adapted biomedical LMs (Luo et al., 2022; Christophe et al., 2024; Google DeepMind, 2025) extend this line. A separate thread couples LLMs with retrieval: retrieval-augmented generation is increasingly used in biomedical question answering (Xiong et al., 2024), typically grounding LLMs in unstructured literature rather than curated ontologies. DeepRare (Zhao et al., 2026) combines LLM reasoning with more than 40 external tools and over 10 biomedical knowledge sources, achieving state-of-the-art performance on several rare-disease benchmarks.

Most closely related to our work are systems that combine an HPO-based phenotype retriever with downstream refinement, but to our knowledge no prior work has systematically compared retriever-only, retrieve-and-rerank, and unrestricted LLM di-

agnosis across a comparable range of benchmarks, or characterized *when* each approach has the advantage and why.

3 Methods

3.1 Task and Metrics

Given a clinical vignette with HPO-encoded phenotypes rendered as natural language, produce a ranked list of OMIM disease diagnoses. We report Recall@k (R@k) and MRR under **strict OMIM ID matching**. Statistical significance: paired McNemar's test; uncertainty: bootstrap 95% CIs (10,000 resamples). Code, prompts, and per-case outputs are released at https://github.com/mofty8/BioNLP26.

We define **coverage** as the fraction of cases where the truth disease appears in PhenoDP's top-50 retrieved candidates. This is a property of each case—a case is *covered* if PhenoDP ranks its true diagnosis within the top 50, regardless of how many HPO terms the patient presents with. Cases where truth is absent from the top 50 are **non-retrievable**—they represent a hard ceiling for any retrieval-based approach that uses the same candidate pool. We measure coverage using PhenoDP as a concrete retriever, but the failure modes we document (§5.1) are properties of HPO/HPOA-anchored retrieval in general and are not specific to PhenoDP's scoring function.

3.2 Retriever: PhenoDP

PhenoDP (Wen et al., 2025) is a deep-learning-based toolkit whose Ranker module scores candidate diseases by combining three similarity measures between a patient's HPO terms and each disease's HPOA profile: information-content (IC) similarity, phi-coefficient similarity after graph propagation, and semantic similarity via contrastive embeddings. Wen et al. (2025) report that IC dominates, phi contributes, and semantic similarity is minimal; in practice PhenoDP's ranking is driven primarily by term-level information content. The Ranker searches the full HPOA knowledge base (12,717 annotated diseases) and returns up to 50 candidates per case; its performance is therefore bounded by HPOA completeness—a limitation central to our analysis.

3.3 Reranking and Unrestricted Diagnosis

We evaluate the LLM under two conditions that share an input format but differ in whether a retriever is present.

Retrieve-and-rerank. The LLM reranker receives (i) a clinical narrative generated by PP2Prompt (Reuter et al., 2024), which converts HPO terms into natural language, and (ii) PhenoDP's top-k candidates ($k \in \{3, 5, 10\}$). Unless otherwise noted, we report top-10 results as this consistently yielded the best performance. We evaluate two prompt strategies: a *rule-gated* prompt with explicit conditional rules (preserve the retriever's ordering when the top-1 score exceeds rank-2 by a large margin, unless specific phenotypic contradictions are identified; full template in Appendix D), and a *narrative* prompt adapted from PP2Prompt (Reuter et al., 2024) that presents candidates without conditional gates and allows free reordering.

Unrestricted LLM diagnosis. The LLM receives only the clinical narrative and produces up to 10 ranked diagnoses from parametric knowledge, with no retriever input.

3.4 Models

For reranking: Gemma-3 27B (Gemma Team, 2025), Gemma-4 31B,[1] Llama-3.3 70B (Meta AI, 2024). For unrestricted diagnosis: additionally Med42-70B and Med42-8B (Christophe et al., 2024), MedGemma-27B (Google DeepMind, 2025), and Qwen2.5-32B (Qwen Team, 2024); Table 2 shows the two strongest open-LLM baselines, and full results for all seven models are in Appendix A. All open-weight, instruction-tuned, temperature 0.

3.5 Datasets

We evaluate on seven benchmarks (Table 1): Phenopackets (Wen et al., 2025), four RareBench subsets (Chen et al., 2024) (LIRICAL, Matchmaker Exchange [MME], RAMEDIS, Hereditary Monogenic Syndromes [HMS]), MyGene2 (Chong et al., 2020), and DDD[2]—curated disease-level HPO profiles from the Deciphering Developmental Disorders database, the same source used by DeepRare. Our DDD subset (n=2,248) differs slightly

[1] Gemma-4 (https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/); no separate technical report was available at the time of this evaluation, so we cite the official announcement and release the exact checkpoint identifier alongside our code.

[2] https://www.deciphergenomics.org/ddd/ddgenes

Dataset	Type	n	Dis.	Cov (%)	HPOA med
Phenopackets	Lit	6901	500	81.5	47
LIRICAL	Lit	370	411	85.7	43
MME	Clin	40	28	95.0	47
DDD	Curated	2248	2171	79.2	22
MyGene2	CR	131	37	70.2	42
RAMEDIS	CR†	624	160	56.7	31
HMS	Clin	68	101	48.5	50

Table 1: Datasets (10,382 cases total). Shaded rows are *low-coverage* (<60%). **Type** codes: Lit = published case reports with author-provided HPO terms; Clin = prospectively collected hospital records; CR = case reports with HPO terms mapped post hoc; Curated = disease-level HPO profiles from DECIPHER (DDD is not directly comparable to patient-level benchmarks; see §3.5). **Columns**: n evaluable cases; **Dis.** unique truth diseases; **Cov** retrieval coverage (fraction of cases whose truth is in PhenoDP's top-50); **HPOA med** median HPOA annotations for the truth diseases. †RAMEDIS HPO terms were manually mapped from clinical descriptions by Chen et al. (2024). HMS uses the n=68 OMIM-evaluable subset; MyGene2 uses the n=131 PhenoDP-compatible subset.

from DeepRare's (n=2,283) due to filtering for PhenoDP-compatible OMIM diseases; we were unable to exactly reproduce their case selection as the original split is not publicly available. For HMS, the original dataset contains 88 cases; 20 are annotated with ORPHA-only IDs corresponding to non-Mendelian immune disease groupings that do not map to specific OMIM entries and therefore lack the HPOA phenotype annotations required by the retrieval stack, leaving 68 evaluable cases.

4 Results

4.1 Cross-Approach Comparison

Table 2 presents results across the three approaches (retriever alone, retrieve-and-rerank, unrestricted LLM) alongside DeepRare's published numbers. DeepRare results are taken directly from their supplementary table (Zhao et al., 2026); direct numerical comparison is approximate throughout, as DeepRare does not report Phenopackets results and evaluation conditions may differ from ours. Two main patterns emerge.

The reranker improves over retrieval. On Phenopackets, the best reranker (Gemma-4 31B, rule-gated, top-10) improves R@1 from 51.2% to 54.3% (+3.1 pp, McNemar $p < 0.0001$; 262 promotions vs. 67 demotions). On MME, +5.0 pp (65.0% → 70.0%). On LIRICAL, +1.4 pp. On RAMEDIS, the narrative prompt achieves the largest gain: +12.9 pp (10.3% → 23.2%, Gemma-3 27B). The reranker does not improve on DDD (59.0% → 59.0%): DDD is a curated disease-level benchmark whose phenotype profiles are drawn from the same ontological sources as HPOA, so the ranking the retriever produces largely reflects alignment between two versions of the same knowledge base (see §3.5 and Limitations). The combined pipeline reaches numbers comparable to DeepRare DS-V3 on MME (70.0% vs. 70.0%, n=40) and is within 2.5 pp of DeepRare GPT-4o (72.5%), and is similarly close on LIRICAL (54.1% vs. 56.0%). We caution that these comparisons are approximate rather than head-to-head: our pipeline uses open-weight models up to 31–70B against DeepRare's GPT-4o / DeepSeek-V3 backbones, DeepRare's case selection and prompting may differ from ours, and the MME and LIRICAL sample sizes (40 and 370) give wide bootstrap CIs (Table 3). The point is that an open-weight two-component pipeline is in the same neighborhood on high-coverage datasets, not that it is statistically equivalent.

The performance crossover across datasets. On MME, retrieval achieves 65.0% versus the unrestricted LLM's 5.0% (a 60.0 pp gap in favor of retrieval). On RAMEDIS, the LLM achieves 46.8% versus the retriever's 10.3% (a 36.5 pp gap in the opposite direction). This crossover is statistically significant on six of seven benchmarks (McNemar $p < 0.0001$; Table 3); HMS (n=68 shared cases) does not reach conventional significance (p=0.052) due to limited statistical power.

4.2 The Regime Shift

Figure 1 visualizes the regime shift from two angles. The scatter plot (a) shows retrieval coverage is strongly associated with which approach dominates, with the datasets in our study separating into two clusters: five at ≥70% coverage and two at ≤57%, with no observations in the 60–70% range. The boundary should therefore be read as a hypothesis consistent with these data rather than a located threshold; identifying its precise location requires benchmarks in the transition zone. The bar chart (b), with datasets ordered by decreasing coverage, makes the crossover visible: reading left to right, the dominant system switches from retrieval (blue) to unrestricted LLM (orange).

		Pheno.		LIRICAL		MME		DDD		MyG2		RAMED.		HMS	
System	**Model**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
Retriever	PhenoDP	51.2	67.2	52.7	70.8	65.0	85.0	59.0	70.2	38.9	61.8	10.3	28.7	7.4	25.0
Reranker	Gem-4 31B	**54.3**	67.6	54.1	75.9	70.0	90.0	59.0	70.2	48.9	61.8	18.3	37.0	14.7	36.8
Reranker	Lla-3.3 70B	52.6	67.3	54.1	75.9	65.0	90.0	**59.1**	70.2	38.9	61.8	22.9	37.0	16.2	36.8
Open LLM	Lla-3.3 70B	14.3	25.6	17.0	28.4	5.0	10.0	12.6	23.1	3.8	15.3	46.0	66.7	17.6	30.9
Open LLM	Med42 70B	11.5	24.1	15.9	26.2	2.5	2.5	6.4	12.8	2.3	16.0	46.8	64.1	20.6	27.9
DeepRare	DS-V3	—[a]	—	56.0	—	70.0	—	43.0	—	72.6	—	72.6	—	**57.0**	—
DeepRare	GPT-4o	—	—	**56.0**	—	**72.5**	—	43.0	—	**75.9**	—	**73.2**	—	54.5	—

Table 2: Cross-approach comparison. Shaded rows: our pipeline. **Bold**: best R@1 per dataset across all systems. The reranker rows report the best (prompt, top-k) configuration on each dataset's test set (see Table 4 and Limitations); fixed-configuration sensitivity is reported for Phenopackets in Appendix B. Green rows: DeepRare from Zhao et al. (2026); DeepRare results are not available for Phenopackets ([a]) and direct comparison is approximate for all datasets (§3.5). All our metrics use strict OMIM ID matching. All HMS numbers use the n=68 OMIM-evaluable subset; all MyGene2 numbers use the n=131 PhenoDP-compatible subset (§3.5). (—): R@5 not available for DeepRare.

Dataset	**Ret**	**LLM**	**CI$_{\text{Ret}}$**	p	**Setting**
Pheno.	51.2	11.5	50.0–52.4	<.001	Retrieval
LIRICAL	52.7	15.9	47.6–57.8	<.001	Retrieval
MME	65.0	2.5	50.0–80.0	<.001	Retrieval
DDD	59.0	12.6	57.7–60.3	<.001	Retrieval
MyGene2	38.9	3.8	30.7–47.5	<.001	Retrieval
RAMEDIS	10.3	46.8	8.0–12.7	<.001	LLM
HMS	7.4	20.6	1.5–14.7	.052	—[†]

Table 3: Statistical significance of the retrieval vs. unrestricted LLM comparison (McNemar test on cases shared between both systems; LLM = Med42 70B). CI: bootstrap 95% for retriever R@1. MyGene2 numbers use the n=131 PhenoDP-compatible subset. [†]HMS: n=68, underpowered.

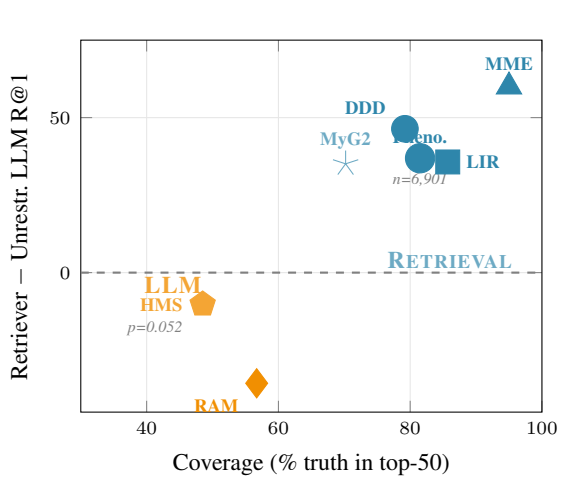
Dataset	**Config**	**Ret**	**Rer**	**Δ**	**P**	**D**
Retrieval regime						
Pheno.	G4/Rule/10	51.2	**54.3**	+3.1	262	67
LIRICAL	G4/Narr/10	52.7	**54.1**	+1.4	10	5
MME	G4/Narr/10	65.0	**70.0**	+5.0	4	2
MyGene2	G4/Narr/10	38.9	**48.9**	+9.9	13	0
LLM regime						
RAMEDIS	G3/Narr/10	10.3	**23.2**	+12.9	78	25
HMS	G3/Narr/10	7.4	**16.2**	+8.8	7	0

Table 4: Reranker performance by dataset, reporting the *best (model, prompt, top-k) configuration on each dataset's test set*; these numbers therefore upper-bound what a single fixed configuration would obtain and should not be read as a held-out estimate (see Limitations and Appendix B for the full Phenopackets grid). **Config** format: Model/Prompt/top-k, where G4=Gemma-4 31B, G3=Gemma-3 27B, L3=Llama-3.3 70B; Rule=rule-gated prompt, Narr=narrative prompt; 10=top-10 candidates. **Ret/Rer**: retriever/reranker R@1 (%). **P**: promotions (reranker moves truth to a higher rank). **D**: demotions (reranker moves truth to a lower rank). DDD is excluded: all configurations yield Δ=0.0 pp. The narrative prompt dominates in the LLM-dominant setting; the rule-gated prompt is optimal in the retrieval-dominant setting (Phenopackets).

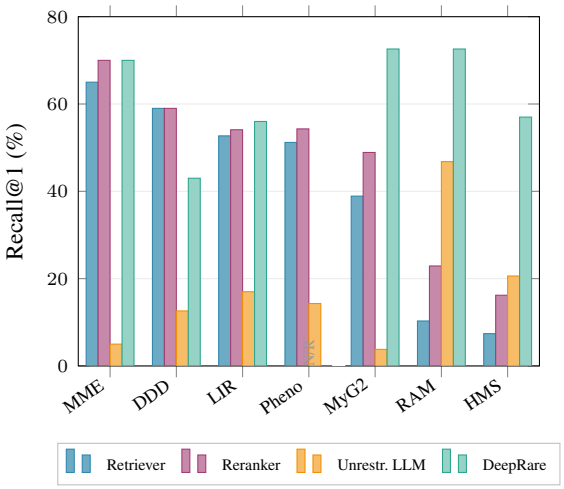
4.3 Reranker Analysis: Per-Subset and Conditional

Table 4 disaggregates the reranker analysis by dataset (DDD excluded, as discussed in §4). First, the reranker improves R@1 on every dataset, with gains ranging from +1.4 pp (LIRICAL) to +12.9 pp (RAMEDIS). On Phenopackets, the 3.9:1 promotion-to-demotion ratio is driven by clean cases like neurofibromatosis type 1 (204 in-window cases: 103 promotions to rank 1, zero demotions); the most-harmed disease is Aarskog-Scott syndrome (15 promotions against 48 demotions), which is consistently confused with Noonan syndrome—the two share facial dysmorphism and short stature but have distinct causal genes. Second, prompt strategy interacts with the dominant approach: the rule-gated prompt excels on Phenopackets (where retrieval is strong and conservative reranking avoids harmful demotions), while the narrative prompt is necessary on RAMEDIS, HMS, and MyGene2 (where the retriever has more room for improvement). Third, the rule-gated prompt is the strongest configuration only on Phenopackets, where retriever scores are well-separated and its conditional rules are triggered; on the shorter RareBench narratives the narrative prompt is chosen as best in Table 4.

Conditional performance on retrievable cases. The overall R@1 numbers in Table 2 mix two effects: how often the retriever places the correct answer in its top-k window (coverage), and how well



(a) Coverage is strongly associated with the dominant approach. All datasets with >70% coverage fall in the retrieval-dominant setting; both with <60% fall in the LLM-dominant setting. Unrestricted LLM: Llama-3.3 70B.



(b) R@1 by approach (decreasing coverage left→right). The crossover is visible. DeepRare adds most value in the LLM-dominant setting (right). Phenopackets not reported by DeepRare.

Figure 1: The diagnostic regime shift. (a) In our seven benchmarks, coverage is associated with which approach dominates; observations cluster at ≥70% or ≤57%, with no data in the 60–70% transition zone, so the crossover boundary is a hypothesis consistent with the data rather than a located threshold. (b) Grouped bars reveal the crossover between approaches.

the reranker reorders candidates once the correct answer is present. To isolate the reranker's discrimination ability, we report a *conditional* ("fair") evaluation restricted to cases where the truth disease is within the retriever's top-k—i.e., cases the retriever could in principle solve and that the reranker actually sees. On RareBench (Llama-3.3, narrative, top-10; LIRICAL + MME + RAMEDIS + HMS, 1,102 cases total), the reranker achieves 65.4% R@1 on the 573 cases whose truth is within the top-10 window, versus the retriever's 50.6% on the same subset (+14.8 pp). On Phenopackets (Gemma-4, rule-gated, top-10), 75.5% versus 71.2% (+4.3 pp). These conditional gains are substantially larger than the overall gains in Table 2, confirming that the LLM adds real clinical reasoning value when the correct answer is already in the candidate set—the modest overall gains are partly an artifact of the hard retrieval ceiling imposed by non-retrievable cases.

5 Explaining the Regime Shift

5.1 Non-Retrievable Cases

The regime shift is directly linked to *non-retrievability*. Table 5 shows the mapping is sharp: the two LLM-regime datasets (RAMEDIS 43.3%, HMS 51.5%) have non-retrievable rates clearly above the retrieval-regime range (5.0–29.8%), and the ordering by non-retrievable rate matches the ordering by which approach wins.

Dataset	Total	Non-ret	Rate	Regime
MME	40	2	5.0%	Retrieval
LIRICAL	370	53	14.3%	Retrieval
Phenopackets	6,901	1,278	18.5%	Retrieval
MyGene2	131	39	29.8%	Retrieval
RAMEDIS	624	270	43.3%	LLM
HMS	68	35	51.5%	LLM

Table 5: Non-retrievable rates align with the regime. DDD is excluded as a curated disease-level benchmark rather than a patient-case dataset (§3.5).

Analysis of the 1,278 non-retrievable cases in the Phenopackets benchmark reveals two structural failure modes:

Annotation sparsity. Non-retrievable diseases have median 24 HPOA entries versus 52 for retrievable ones. Diseases with <20 annotations have 22–37% non-retrievable rates; those with >40 are retrieved at 88.4%. Example: *congenital adrenal hyperplasia (21-hydroxylase deficiency)*—50 cases, only 12 HPOA annotations, 100% non-retrievable.

Phenotypic homogeneity. Developmental and epileptic encephalopathy (DEE) subtypes share near-identical HPO profiles despite distinct genetics: DEE-4 (430 cases, 48% non-retrievable), DEE-11 (260 cases, 82%). HPO-based retrieval cannot disambiguate them. Homogeneity failures are dis-

Dataset	Ret	LLM	Oracle‡	Δ	LLM-only†
Pheno.	51.2	11.5	**56.7**	+5.5	376
LIRICAL	52.7	15.9	**59.7**	+7.0	26
MME	65.0	2.5	**67.5**	+2.5	1
DDD	59.0	12.6	**61.8**	+2.8	63
MyGene2	38.9	3.8	**41.2**	+2.3	3
RAMEDIS	10.3	46.8	**51.1**	+4.3	255
HMS	7.4	20.6	**26.5**	+5.9	13

Table 6: Complementarity of retrieval and unrestricted LLM (Med42 70B) across all seven benchmarks. ‡**Oracle**: R@1 of a hypothetical perfect router that always selects whichever approach answers each case correctly—an upper bound on hybrid system performance. Δ: gain over the better single system. †**LLM-only**: cases answered correctly by the unrestricted LLM but not by the retriever—i.e., cases where only parametric LLM knowledge suffices. Shaded rows: LLM-dominant datasets. Complementarity holds on every dataset; the largest per-dataset oracle gains are LIRICAL +7.0, Phenopackets +5.5, and HMS +5.9.

sociable from sparsity: Xia-Gibbs syndrome is 86% non-retrievable despite 187 HPOA annotations (Appendix C), because its dense profile overlaps with many other syndromic neurodevelopmental disorders.

At category level, epilepsy/neurological (40.2%) and immunodeficiency (42.0%) are highest risk, driven by sparse annotations and phenotypic overlap. Mitochondrial diseases (1.9%) and RASopathies (1.3%) are rarely missed, owing to distinctive phenotypes and rich annotations (median 113 and 56 HPOA entries, respectively). The same sparse-annotation signature appears on other benchmarks: on HMS, non-retrievable diseases have a mean of 8.0 HPOA entries vs. 23.0 for retrievable ones; on RAMEDIS, 23.9 vs. 36.2.

5.2 Complementarity of Retrieval and LLM Diagnosis

Table 6 quantifies how much could be gained by a hybrid system that combines both approaches. A hypothetical perfect router—one that selects whichever approach answers each case correctly—would improve over the better single system by 2.3–7.0 pp across all seven datasets. The key insight is that the two approaches succeed on largely non-overlapping cases. On Phenopackets, 376 patient cases (5.5%) are answered correctly only by the unrestricted LLM; these are predominantly non-retrievable cases where the true disease is absent from PhenoDP's top-50 and only parametric LLM knowledge can recover the answer. On RAMEDIS, 255 cases (40.9%) fall in this category, consistent with its high non-retrievable rate.

These results motivate hybrid systems that combine retrieval and LLM generation. We evaluate a score-based router in §6, finding that retriever confidence alone is insufficient to close the gap and that input-modality features are needed.

6 Discussion

Where agentic complexity adds value. On low-coverage datasets, DeepRare substantially outperforms our pipeline (41–50 pp on RAMEDIS and HMS), suggesting that richer runtime knowledge access matters most in these settings. On high-coverage patient-case datasets (MME and LIRICAL, with n=40 and n=370), our open-weight two-component pipeline lands in the same neighborhood as DeepRare (DDD is excluded from this comparison; see §3.5). We stop short of calling the pipeline "competitive" in a statistical sense: at these sample sizes, a single case flip on MME shifts R@1 by 2.5 pp, and our pipeline uses 31–70B open-weight rerankers against DeepRare's GPT-4o / DeepSeek-V3 backbones under evaluation conditions we could not exactly reproduce. The interpretable claim is structural: on high-coverage settings the simpler pipeline reaches a similar operating point with a smaller open-weight backbone, while on low-coverage settings agentic complexity is doing real work that retrieval-plus-reranking cannot replace. This still supports a practical design principle—route high-coverage cases through a lightweight pipeline and reserve the more resource-intensive agentic system for the cases that genuinely need it—but the boundary of when each is sufficient should be set on larger benchmarks under matched evaluation conditions.

Implications for evaluation. Aggregate metrics across heterogeneous benchmarks obscure the regime shift. Two systems with the same overall average can have very different per-dataset profiles—strong on high-coverage cases but weak on low-coverage ones, or the reverse—and these profiles have very different clinical implications. We recommend per-dataset reporting with explicit coverage characterization, so that readers can understand which cases each system actually handles well.

Toward hybrid systems: a negative result on *pooled* score-based routing. The oracle analysis (Table 6) shows that retrieval and unrestricted LLM

generation succeed on largely non-overlapping cases, with a perfect router improving 2.3–7.0 pp over the better single system across all seven datasets. We tested whether a *single* pooled threshold over PhenoDP confidence signals can serve as a practical routing rule: route to the retriever if a confidence statistic exceeds a threshold τ, otherwise escalate to the LLM. We evaluated four candidate signals: (i) the rank-1 similarity score, (ii) the margin between rank-1 and rank-2 scores, (iii) the coefficient of variation (CV) across the 50 candidate scores—a high CV indicates that one disease scores substantially above the others, while a low CV suggests the candidates are tightly clustered—and (iv) the product of rank-1 similarity and CV. A pooled threshold sweep on rank-1 similarity (N=10,379) finds τ^*=0.705, recovering +1.2 pp (16% of the oracle gap) in-sample; in a leave-one-dataset-out evaluation, the router *hurts* on five of seven held-out datasets (micro-average: 49.7% vs. 50.1% retriever baseline). The other three signals behave similarly: none improve over the single-score baseline in the leave-one-dataset-out setting. This result speaks only to pooled thresholding; per-dataset thresholds, learned routers, and confidence signals derived from sources other than PhenoDP's own scores were not evaluated.

Why a pooled threshold fails. The failure is structural rather than a tuning issue. RAMEDIS, where the LLM leads by 35 pp (46% vs. 10%), needs $\tau \to \infty$ (always escalate to the LLM), while Phenopackets and DDD, where retrieval leads by ≈37 and ≈46 pp respectively, need $\tau \to 0$ (never escalate). A single pooled threshold calibrated on the mix is miscalibrated for both, and no monotone confidence signal over the retriever's own scores can resolve this tension with a pooled rule. This does not rule out per-dataset thresholding or learned routers, which we did not evaluate here.

What a useful router would need. Effective case-level routing requires a signal that reflects *input modality*—whether a case is expressed as structured HPO terms (favoring retrieval) or free-text clinical narrative (favoring parametric LLM knowledge)—rather than retriever confidence alone. Candidate features include HPO annotation density, clinical concept density, and the ratio of specific to general phenotype terms. A learned router combining these with the PhenoDP confidence signals above could in principle exploit the 2–7 pp oracle gains identified in Table 6; pairing such a router with the reranker studied here is a natural next step.

Limitations

Several aspects of this study warrant caution.

1. **Benchmark composition.** Our strongest results are on literature-derived and curated benchmarks (Phenopackets, LIRICAL, MME, DDD), which tend to be the easier diagnostic settings (Zhao et al., 2026); the patterns we report may not transfer to more complex, real-world clinical presentations with noisy phenotypes and incomplete workups.
2. **Model scope.** We evaluate only open-weight models up to 70B parameters; frontier proprietary models used as rerankers could yield different results.
3. **DeepRare comparison.** The comparison uses DeepRare's published per-dataset numbers, which may have been obtained under evaluation conditions different from ours (e.g., different case selection, metrics, or prompting); cross-paper comparisons should be treated as approximate.
4. **DDD is not a patient-case benchmark.** Our DDD subset (n=2,248) does not exactly match DeepRare's (n=2,283) due to filtering for PhenoDP-compatible OMIM diseases. Beyond this case-selection gap, DDD is a curated disease-level HPO profile benchmark rather than a set of individual patient cases, and its phenotype annotations are drawn from the same ontological sources as HPOA, so any retrieval score on DDD partly reflects alignment between two versions of the same knowledge base and overstates the ceiling for real patient cases.
5. **HMS statistical power.** HMS has only 68 evaluable cases (after excluding the 20 ORPHA-only cases without OMIM/HPOA entries), which limits statistical power; our McNemar test on HMS does not reach conventional significance (p=0.052).
6. **Seven-dataset coverage axis.** The coverage-based account of the regime shift is built on seven datasets. Additional benchmarks—particularly ones with moderate coverage in the 60–70% range that lies in the transition zone—would strengthen the empirical case.

7. **Per-dataset configuration selection.** The reranker numbers we report in Tables 2 and 4 use the best (model, prompt, top-k) configuration selected on each dataset's test set. They therefore upper-bound what a single fixed configuration would achieve and do not constitute a held-out estimate. We report a full Phenopackets configuration grid in Appendix B to characterize sensitivity, but analogous grids on the remaining datasets and leave-one-dataset-out configuration selection were beyond the scope of this study.

8. **Retrospective evaluation.** All benchmarks used here have known ground truth. We do not evaluate how any of these systems perform prospectively in clinical workflows, where the distribution of cases and the cost of errors may differ substantially from benchmark conditions.

Ethics Statement

All datasets used here are publicly available and, where applicable, released under protocols covering research use; we did not collect or contact patients, and no new protected health information was generated.

This work is a retrospective methodological study, not a clinical system: none of the pipelines we evaluate are safe for deployment without prospective validation, human oversight, and local calibration. Rare disease diagnosis has asymmetric costs—missed diagnoses prolong the diagnostic odyssey, incorrect ones can trigger unnecessary interventions—so any deployment should treat system output as a suggestion to a clinician, not an answer, and should abstain on low-confidence cases and on regimes outside the system's demonstrated competence (e.g., the low-coverage regime identified here).

Both components inherit biases from their training data. HPOA is not uniformly complete, and the resulting annotation gaps drive the non-retrievability failures we study, producing structural performance disparities across diseases and, by extension, across the patient groups most affected by them. Parametric LLMs over-represent literature-reported presentations, and our benchmarks skew toward published case reports and curated profiles; the numbers here should not be read as predictive of performance on pediatric, low-resource, or non-Western presentations.

Code, prompts, and per-case outputs are released at https://github.com/mofty8/BioNLP26.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – RTG2424/CompCancer – project number 377984878.

References

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. RareBench: Can LLMs serve as rare diseases specialists? In *Proceedings of KDD*.

Jessica X Chong and 1 others. 2020. MyGene2: improving the utility of undiagnosed disease web resources. *American Journal of Human Genetics*, 106:287–291.

Clément Christophe and 1 others. 2024. Med42-v2: A suite of clinical LLMs. *arXiv preprint arXiv:2408.06142*.

Toyofumi Fujiwara, Yasunori Yamamoto, Jin-Dong Kim, Orion Buske, and Toshihisa Takagi. 2018. PubCaseFinder: a case-report-based, phenotype-driven differential-diagnosis system for rare diseases. *The American Journal of Human Genetics*, 103:389–399.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2025. MedGemma: Medical AI models from Google. *arXiv preprint*.

Melissa Haendel, Nicole Vasilevsky, and 1 others. 2020. How many rare diseases are there? *Nature Reviews Drug Discovery*, 19:77–78.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Sebastian Köhler and 1 others. 2021. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217.

Renqian Luo and 1 others. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Harsha Nori and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? *arXiv preprint arXiv:2311.16452*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Max S Reuter and 1 others. 2024. Evaluation of LLMs for rare disease differential diagnosis using structured phenotype data. *medRxiv*.

Peter N Robinson, Vida Ravanmehr, Michael A Gargano, and 1 others. 2020. LIRICAL: an approach to prioritize candidate diseases for rare disease diagnostics. *American Journal of Human Genetics*, 107:168–176.

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371:2039–2041.

Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. Identifying and extracting rare diseases from clinical notes using generative large language models. *Journal of the American Medical Informatics Association*, 31(4):909–918.

Karan Singhal and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Damian Smedley, Julius OB Jacobsen, Maximilian Jäger, and 1 others. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10:2004–2015.

Stephanie Nguengang Wakap and 1 others. 2020. Estimating cumulative point prevalence of rare diseases. *European Journal of Human Genetics*, 28:165–173.

Baole Wen, Sheng Shi, Yi Long, Yanan Dang, and Weidong Tian. 2025. PhenoDP: leveraging deep learning for phenotype-based case reporting, disease ranking, and symptom recommendation. *Genome Medicine*, 17:67.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Weike Zhao, Chaoyi Wu, Yanjie Fan, and 1 others. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651:775–784.

A Full Unrestricted LLM Comparison

B Full Reranker Grid: Phenopackets

C Non-Retrievable Case Studies

The two structural failure modes identified in §5.1 (annotation sparsity and phenotypic homogeneity) are best understood through concrete examples.

Model	R@1	R@5	R@10	MRR
Llama-3.3 70B	14.3	25.6	29.9	.192
Gemma-4 31B	12.2	19.9	23.1	.156
Med42-70B	11.5	24.1	27.8	.171
Gemma-3 27B	11.3	17.4	20.0	.140
Qwen2.5-32B	10.5	17.1	19.5	.133
MedGemma 27B	10.2	15.8	17.3	.126
Med42-8B	7.1	10.3	11.4	.085

Table 7: Unrestricted LLMs on Phenopackets (strict OMIM ID).

Model	Prompt	K	R@1	Δ	MRR
Gemma-4	Rule	10	**54.3**	+3.1	**.601**
Gemma-4	Rule	5	53.7	+2.5	.590
Gemma-4	Rule	3	53.5	+2.2	.578
Llama-3.3	Rule	10	52.6	+1.4	.589
Llama-3.3	Rule	5	52.3	+1.1	.581
Llama-3.3	Rule	3	52.4	+1.2	.572
Gemma-3	Rule	10	51.6	+0.4	.584
Llama-3.3	Narr	3	51.8	+0.6	.568
Gemma-4	Narr	10	50.7	−0.6	.575
Gemma-3	Narr	10	48.8	−2.5	.565

Table 8: Phenopackets reranker grid (n=6,901, PhenoDP baseline: 51.2%). The rule-gated prompt consistently improves; the narrative prompt can harm on this high-coverage dataset.

Annotation sparsity: congenital adrenal hyperplasia (21-hydroxylase deficiency). This is the single most striking failure in the Phenopackets benchmark: 50 patient cases, all 50 non-retrievable (100%). The disease has only 12 HPO annotations in HPOA—a fraction of the 52-term median for retrievable diseases. Patient phenotype sets typically include electrolyte disturbance, virilization, ambiguous genitalia, salt wasting, and adrenal crisis, but the sparse HPOA profile does not encode enough of these terms for the information-content-based similarity used by PhenoDP to pull the disease into the top-50. An unrestricted LLM, by contrast, recognizes the classic clinical pattern from training text and returns the correct diagnosis immediately. This is a clean illustration of why coverage, not reasoning, is the bottleneck in such cases: no amount of LLM reranking can rescue a candidate that is never retrieved.

Phenotypic homogeneity: the DEE cluster. Developmental and epileptic encephalopathies (DEE-4, DEE-9, DEE-11, and related subtypes) share nearly identical HPO profiles despite distinct underlying genes (KCNQ2, SCN1A, STXBP1, and others). DEE-4 accounts for 430 cases in Phenopackets with a 48% non-retrievable rate; DEE-11 ac-

counts for 260 cases with an 82% non-retrievable rate. When truth is in the top-50, the retriever ranks several phenotypically indistinguishable subtypes together, and the information needed to disambiguate—gene panels, age at onset, EEG pattern, response to specific anti-epileptic drugs—is not encoded in the HPO terms the retriever sees. This is the homogeneity failure mode: retrieval has no mechanism to break ties between entities that look the same in ontology space. Xia-Gibbs syndrome (65 cases, 86% non-retrievable despite 187 HPOA annotations) fails for the same reason at a broader scale—its phenotype profile is dense but overlaps with many other syndromic neurodevelopmental disorders.

Category-level pattern. These case-level failures aggregate into category-level signatures (§5.1). Immunodeficiency (42.0%) and epilepsy/neurological (40.2%) categories are dominated by sparse-annotation or phenotypically overlapping disease families and account for most non-retrievable cases. Mitochondrial diseases (1.9%) and RASopathies (1.3%) are rarely missed—both have distinctive multi-system phenotypes and dense annotations (median 113 and 56 HPOA entries, respectively). The practical implication is that non-retrievability is predictable from disease identity, not random; a routing policy that escalates cases in the high-risk categories to an LLM is therefore a sensible starting point for hybrid systems (§6).

D Prompt Templates

For reproducibility, we include the literal prompt templates used for the rule-gated and narrative reranking conditions. Both prompts consume (i) the PP2Prompt-generated clinical narrative and (ii) PhenoDP's top-k candidates with similarity scores; models are asked to return a ranked list of k OMIM IDs (inference settings are described in §3). The rule-gated prompt enforces three conservative gates (score-gap, obligatory-feature, default-to-retriever) and changes the ranking in $<5\%$ of cases on Phenopackets. The narrative prompt, adapted from PP2Prompt (Reuter et al., 2024), omits the gates and produces 3–5$\times$ more rank changes; it is beneficial on low-coverage datasets but can harm high-coverage ones (e.g., DDD: -4.7 pp).

Rule-gated prompt.

```
You are an expert clinical geneticist.
You will rerank candidate rare diseases
for a patient. Apply the following rules
in order before changing the ranking.

[Score-gap gate]
- If rank-1 score exceeds rank-2 by >= 0.15,
  preserve the retriever's ordering unless
  a Condition A or Condition B violation
  is identified.
- If rank-1 score exceeds rank-2 by >= 0.03
  but < 0.15, you may swap rank-1 and rank-2
  only if a Condition A or Condition B
  violation is identified for rank-1.

[Condition A: obligatory absent feature]
Demote a candidate if ALL four hold:
  (1) the candidate has an obligatory
      HPO feature (frequency "always"
      or "very frequent") in HPOA,
  (2) that feature is in a clinically
      testable category for this patient
      (not age- or assessment-gated),
  (3) the patient record explicitly lists
      the feature as absent (negated),
  (4) absence is not trivially explained
      by the patient's age or workup stage.

[Condition B: biological impossibility]
Demote a candidate if the patient has a
finding that is biologically incompatible
with the candidate disease (e.g., male
patient with an X-linked female-specific
disorder).

[Same-gene spectrum rule]
When multiple candidates differ only in
severity labels for the same gene (e.g.,
"mild" vs. "severe" forms of the same
allelic series), preserve the retriever's
order unless a clear severity marker in
the patient record favors a different form.

[Numbered-subtype rule]
When candidates are numbered subtypes of
the same disease family (e.g., DEE-4 vs.
DEE-11), do not reorder unless a gene,
biomarker, or distinctive phenotype is
explicitly named in the patient record.

Patient narrative:
{narrative}

Candidate diseases (top-{k}):
{candidates_with_scores}

Output: a ranked list of {k} OMIM IDs,
one per line, most likely first. Include
a one-sentence justification for every
change you make relative to the retriever
order, referencing specific HPO terms or
conditions above.
```

Narrative prompt.

```
You are an expert clinical geneticist.
Read the patient narrative carefully and
rank the candidate rare diseases from
most to least likely.

Patient narrative:
{narrative}

Candidate diseases (top-{k}):
{candidates_with_scores}

Produce a ranked list of {k} OMIM IDs,
one per line, most likely first. Base your
ranking on clinical judgment: consider
the full phenotype picture, obligatory
features, and which candidate best
explains the patient. Provide a brief
justification for your top pick.
```

The full prompt strings—including system messages, few-shot examples where used, and the exact candidate-list formatting—are released alongside the code.

BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs

Nourah M. Salem
University of Colorado Anschutz Medical Campus, USA
University of Chicago, USA
nourah(dot)salem@cuanschutz.edu

Elizabeth White and Michael Bada and Lawrence Hunter
University of Chicago, USA

Abstract

Coreference resolution in biomedical texts presents unique challenges due to complex domain-specific terminology, high ambiguity in mention forms, and long-distance dependencies between coreferring expressions. In this work, we present a comprehensive evaluation of generative large language models (LLMs) for coreference resolution in the biomedical domain. Using the CRAFT corpus as our benchmark, we assess the LLMs' performance with four prompting experiments that vary in their use of local, contextual enrichment, and domain-specific cues such as abbreviations and entity dictionaries. We benchmark these approaches against a discriminative span-based encoder, SpanBERT, to compare the efficacy of generative versus discriminative methods. Our results demonstrate that while LLMs exhibit strong surface-level coreference capabilities, especially when supplemented with domain-grounding prompts, their performance remains sensitive to long-range context and mentions ambiguity. Notably, the LLaMA 8B and 17B models show superior precision and F1 scores under entity-augmented prompting, highlighting the potential of lightweight prompt engineering for enhancing LLM utility in biomedical NLP tasks.

1 Introduction

Coreference resolution is the process of identifying entities mentioned in text and grouping all mentions that refer to the same underlying entity (Liu et al., 2023). In the biomedical domain, coreference resolution is a particularly difficult task and has been identified as a key bottleneck in biomedical text mining (Lu and Poesio, 2021). Research articles in this domain often contain dense, technical language, frequent use of abbreviations, and complex referential expressions that rely on domain-specific background knowledge. For instance, resolving a phrase like "the same strain" in a methods section may require linking it back to the "C57BL/6J mice" mentioned several paragraphs earlier, with no intervening repetition or synonyms. Similarly, phrases such as "the compound" may ambiguously refer to any of several chemical entities introduced earlier in experimental descriptions, particularly when multiple drugs or treatments are discussed in parallel. In such cases, surface string similarity offers little guidance; instead, linguistic disambiguation must be informed by contextual and semantic cues.

Adding to the challenge, many biomedical entities share identical surface forms, e.g., a gene and its corresponding protein often have the same name or abbreviation which can confuse automated systems. When clustered by identical surface strings, approximately 65% of the coreference clusters in CRAFT corpus (Cohen et al., 2017b) consist of repeated mentions (Li et al., 2022), emphasizing the need for models that can handle referential ambiguity. Moreover, many coreference links span large textual distances, exceeding the effective context window of conventional models (Lu and Poesio, 2021; Li et al., 2022). These long-range dependencies and requirements for specialized knowledge contribute to the poor generalization of general-domain coreference systems in biomedical contexts.

Given the emergence of increasingly capable large language models (LLMs), a natural question arises: how well can these general purpose models perform coreference resolution in specialized domains like biomedicine, without any task-specific fine-tuning (Gan et al., 2024)? LLMs have demonstrated remarkable abilities in complex reasoning and language understanding via prompt-based zero-shot or few-shot learning, often surpassing traditional models in general-domain tasks. This raises the possibility that their extensive pretraining enables them to handle intricate referential structures, even in domain-specific contexts. While biomedi-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 519–530
July 3-4, 2026 ©2026 Association for Computational Linguistics

cal coreference remains a demanding task, recent advances suggest that with well-designed prompts and minimal scaffolding, LLMs may be more capable than previously assumed.

In this work, we evaluate coreference resolution in biomedical texts using two contrasting approaches: a span-based model (SpanBERT-Large) (Joshi et al., 2020) trained on general-domain data, and several generative LLMs (LLaMA series) (Touvron et al., 2023) prompted to resolve coreference without fine-tuning. Our experiments use the CRAFT corpus, a richly annotated biomedical dataset.

The contributions and objectives of this paper are summarized as follows:

- **Benchmarking open-weight LLMs** We compare three LLaMA models under different prompting strategies: local-only, contextual, abbreviation-aware, and entity-aware against a span-based baseline, reporting performance on the CRAFT corpus.
- **Domain Analysis:** We identify coreference challenges unique to biomedical text, such as identical mention strings and abbreviation ambiguity, and analyze how each model type handles them through qualitative examples and error patterns.

2 Related Work

Coreference resolution in biomedical text is a particularly challenging task due to complex domain-specific terminology, high referential ambiguity, and long-range dependencies. Traditional span-based models such as the end-to-end neural coreference models such as SpanBERT (Joshi et al., 2020) have demonstrated strong performance in general domains. However, their reliance on limited context windows and the need for supervised fine-tuning limits their applicability in biomedical settings, where coreference often requires broader semantic grounding.

Traditional approaches also include rule-based and statistical systems (O'Connor and Heilman, 2013; Ng and Cardie, 2002; Soon et al., 2001), followed by neural architectures such as the mention-ranking model (Clark and Manning, 2016) and end-to-end span-ranking networks (Lee et al., 2017; Durrett and Klein, 2013; Wiseman et al., 2015; Lee et al., 2018). SpanBERT (Joshi et al., 2020) further improved coreference resolution by introducing span-centric pretraining objectives, achieving state-of-the-art results on the OntoNotes (Dobrovolskii, 2021) benchmark. Despite these advances, such models rely heavily on supervised training and domain-specific tuning, limiting their generalizability to out-of-domain settings like biomedical text.

Large language models (LLMs) like OpenAI GPT (Radford et al., 2018) and LLaMA have demonstrated strong zero-shot capabilities in various NLP tasks (Brown et al., 2020; Touvron et al., 2023), including aspects of coreference. Recent studies evaluated LLMs' abilities on pronoun resolution and Winograd schemas (Liu et al., 2024; Wu et al., 2020) for other downstream tasks such as question-answering and query-based span prediction problems. However, few studies have directly assessed LLMs on span-based or noun phrase coreference, particularly in long or technical documents. Most relevant to our work are recent prompting frameworks that use generative models for structured information extraction (Xie et al., 2022), though coreference-specific prompting remains underexplored, especially in specialized domains like biomedical literature.

3 Methods and Materials

3.1 Data

The Colorado Richly Annotated Full-Text (CRAFT) corpus (Cohen et al., 2017a) is a biomedical dataset of 67 full-text, open-access journal articles drawn from PubMed Central, with extensive manual annotations for biomedical concepts, syntactic structure, and coreference identity chains, In Version 2.0[1]. For our experiments, we evaluate language models on 50 articles selected at random, which contained more than 80K entities coreferenced to test the LLMs one.[2].

Three stacking properties make CRAFT uniquely demanding for large language models. First, its coreference chains stretch across *full* journal articles rather than short excerpts: as shown in Figure 1, while 76.14% of coreferential links fall within 500 words, over 23% extend up to 12,000 words, well beyond the effective context window of most conventional neural architectures. Second, roughly 65% of its coreference clusters consist of mentions with identical surface strings, for example, a gene and its cognate protein sharing a name or abbreviation, forcing models to disambiguate

[1] https://github.com/UCDenver-ccp/CRAFT
[2] https://github.com/biocoref

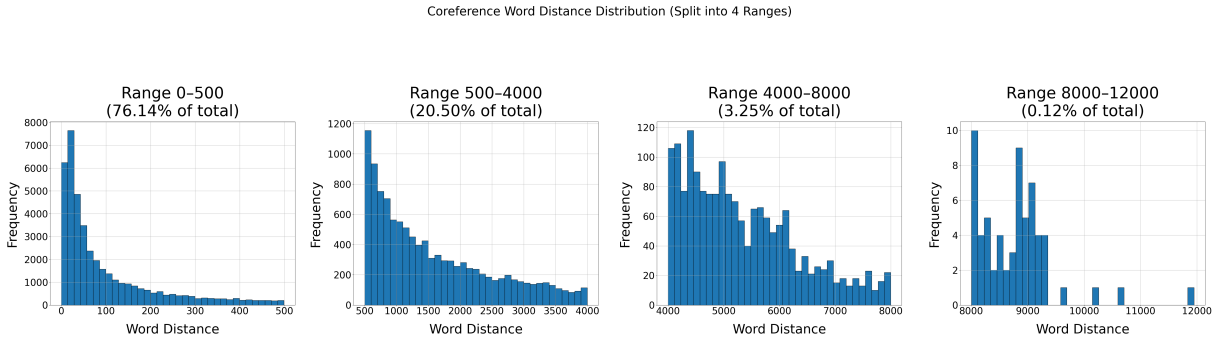


Figure 1: Distribution of word distances between coreferent mentions in biomedical texts, grouped into four ranges.

on semantics rather than lexical cues. Third, CRAFT's dense domain-specific terminology and pervasive abbreviation use cause general-domain systems to collapse: an out-of-the-box coreference resolver achieves only $F_1 = 0.14$ (B^3) on CRAFT, and even a domain-adapted rule-based system reaches only $F_1 = 0.42$ (Cohen et al., 2017a). These properties set CRAFT apart from otherwise strong alternatives. The closest runner-up for hard coreference evaluation is LitBank (Bamman et al., 2020), a challenging long-document complement to OntoNotes (Pradhan et al., 2013) that contains 210,532 tokens across 100 works of English fiction, with an average document length of 2,105 words, roughly four times longer than OntoNotes' 467-token average, and whose narratives form the backbone of recent LLM-oriented coreference benchmarks such as IdentifyMe (Manikantan et al., 2024). However, LitBank truncates each work to its first 2,000 tokens, covers only six ACE-style entity categories, and consists of general-domain prose. CRAFT, by contrast, pairs full-article context with the identical-surface ambiguity and specialized-knowledge demands described above, and couples its coreference layer to eight ontology-grounded semantic classes that yield a 76% increase in named-entity coverage through IDENTITY chains (Cohen et al., 2017a). These factors together establish CRAFT as a strictly harder benchmark than LitBank for evaluating LLM coreference capabilities.

3.2 Models

For our span-based baseline, we evaluate **SpanBERT-large-cased** model (Joshi et al., 2020), a span-optimized transformer pretrained on masked span prediction, on one experiment of coreferencing resolution. Input documents are segmented into 150-word chunks using `Stanza` (Qi et al., 2020), and 500 words in a another experiment setting. We normalized the document chunking sizes using stanza for all the experiments to assure the same chunk indices production for proper evaluation. After chunking the document, noun/pronoun mentions are extracted via `spaCy` (Vasiliev, 2020). Each mention is encoded using SpanBERT's final-layer embeddings and clustered via agglomerative clustering with cosine similarity ($\tau = 0.4$) to group together mentions that semantically refer to the same entity, approximating coreference. For the generative approach, we evaluate three open-weight LLaMA models on each of the 4 coreferencing experiments:

- **LLaMA 3.3 70B-Instruct** (Meta AI, 2024b): a high-capacity model (128k context) released in April 2024.
- **Llama-3.1-8B-Instruct** (Meta AI, 2024a): a compact model optimized for efficient text-only inference.
- **LLaMA 4 Scout 17B** (Meta AI, 2025): a 2025 multimodal model with a 10M-token context window and Mixture-of-Experts architecture.

4 Experiments

We evaluate four prompt-based strategies for coreference resolution using large language models over CRAFT-formatted biomedical texts. Each document D is split into paragraphs $p_1, \ldots, p_N$, each containing approximately 200 words. Using Stanza, we segment the text into sentences and iteratively append them to each paragraph chunk until the 200-word threshold is reached. If the last sentence causes the word count to exceed 200, it is deferred to the next paragraph. The goal is to

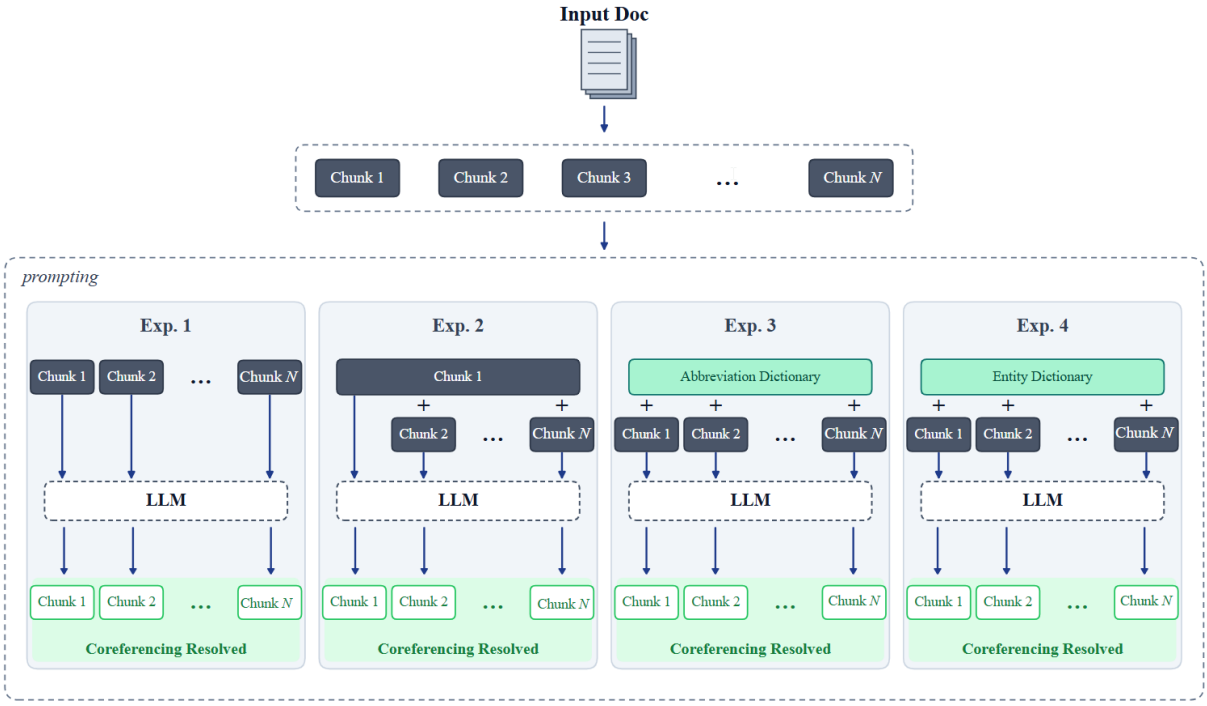


Figure 2: Overview of the coreference resolution pipeline under four prompting strategies. Each chunk is processed by an LLM independently (Exp. 1), with prior context (Exp. 2), or with auxiliary inputs such as abbreviation (Exp. 3) or entity dictionaries (Exp. 4).

output the set of detected mentions M_i, their corresponding resolutions A_i, and a "resolved" version of each paragraph R_i, where each p_i is independently rewritten. Formally:

- Let $\text{LLM}_\phi(\cdot)$ denote the output of the LLM with prompt ϕ.
- Let M_i be the set of coreferent mentions detected in paragraph p_i.
- Let A_i be the set of antecedent resolutions for M_i.
- Let R_i be the rewritten paragraph p_i with all mentions in M_i resolved using A_i.
- The reconstructed document is $\hat{D} = [R_1, \ldots, R_N]$.

The coreference resolution task involves resolving 4 categories: pronouns, definite and indefinite noun phrases, and abbreviations, as illustrated in Table 1.

To evaluate how different categories of auxiliary information affect coreference resolution, we design four prompting configurations: (1) a local-only setup with no external context, (2) a reference-based setup that incorporates the first paragraph as a fixed disambiguation source, (3) an abbreviation-aware setup using a dictionary of extracted abbreviation-definition pairs, and (4) an entity-aware setup using a list of biomedical entities extracted from the document. Algorithm 1 summarizes these 4 styles of the prompting experiments.

Coreference Type	Example Expressions
Pronouns	*it*, *these*, *those*, *its*, *their*
Definite noun phrases	*the gene*, *such results*
Indefinite noun phrases	*a protein*, *some genes*, *one of the enzymes*
Abbreviations	*IOP* $\rightarrow$ *intraocular pressure*

Table 1: Coreference categories and example expressions used in our experiments.

Figure 2 provides an overview of the experimental pipeline used to assess the effectiveness of LLMs across these different coreference categories.

4.1 Experiment 1: Local-Only Resolution (Baseline)

$$R_i = \text{LLM}_{\text{local}}(p_i)$$

In this initial experiment, we investigate the effectiveness of local coreference resolution by prompting LLMs to resolve coreference chains within short, isolated 200-word segments of a biomedical article. Each chunk is independently passed to the LLM. The goal is to assess how well the model performs coreference resolution without any cross-paragraph or global context.

This design reflects a naïve but computationally inexpensive strategy: it minimizes prompt complexity and token limits, while simulating how local context alone may or may not suffice for resolving biomedical coreference phenomena. We made a separate inference run for the 4 coreferencing categories.

This framework allows us to isolate and quantify the limitations of local-only resolution in biomedical texts. It also establishes a baseline against which we can measure subsequent experiments incorporating other resources, such as abbreviation expansion (Experiment 3).

4.2 Experiment 2: Coreference Resolution with Local and Reference Context

$$R_i = \text{LLM}_{\text{ref}}(p_1, p_i)$$

- **Prompt:** Provide p_1 and p_i, instructing the LLM to use the former to disambiguate references in the latter.
- **Purpose:** Test the incremental benefit of a reference paragraph for resolving inter-sentential and cross-paragraph coreferring mentions, without overloading the prompt size.

Building upon the limitations identified in Experiment 1, where coreference resolution was performed in isolation within 200-word chunks, we introduce an additional layer of local context to guide the LLM. In this experiment, each prompt to the model includes not only the target paragraph, but also the first 200-word paragraph in the paper, which carries most of the referential information introduced in the paper and can therefore act as an answer key for the unresolved references in the target paragraph.

Each prompt is structured with two segments: a reference block (Paragraph 1) and a focus block (Paragraph n), with explicit instructions for the LLM to resolve all ambiguous mentions in the focus block using context from the reference. This experiment assesses the impact of lightweight contextual bridging on coreference resolution quality. Compared to the purely local setting in Experiment 1, this approach tests whether even a single paragraph of surrounding context can significantly improve the coherence and referential clarity of LLM-generated outputs, without exceeding typical token limits or requiring full-document inputs.

Algorithm 1 Prompt-Based Coreference Resolution

Require: Document D, ExperimentType $\in$ {LOCAL, REF_CTX, ABBR, ENTITY}, Model LLM
Ensure: ResolvedDocument $\hat{D}$, MentionSets $\mathcal{M}$, ResolutionSets $\mathcal{A}$

Split D into paragraphs: $[p_1, p_2, \ldots, p_N]$
Initialize auxiliary content $C \leftarrow \emptyset$
if ExperimentType = ABBR **or** ExperimentType = ENTITY **then**
  $C \leftarrow$ EXTRACTCONTEXTINFO(D, type=ExperimentType)
end if
Initialize $\hat{D} \leftarrow$ [ ], $\mathcal{M} \leftarrow$ [ ], $\mathcal{A} \leftarrow$ [ ]
for $i = 1$ to N **do**
  $p \leftarrow p_i$
  if ExperimentType = REF_CTX **then**
    reference $\leftarrow p_1$ {Use Paragraph 1 as fixed reference}
  else
    reference $\leftarrow \emptyset$
  end if
  prompt $\leftarrow$ BUILDPROMPT(reference, p, C, ExperimentType)
  response $\leftarrow$ QUERYLLM(LLM, prompt)
  result $\leftarrow$ PARSEJSON(response)
  $\hat{D}$.append(result["Rewritten_Paragraph"])
  $\mathcal{M}$.append(result["Extracted_Expressions"])
  $\mathcal{A}$.append(result["Resolutions"])
end for
return $\hat{D}$, $\mathcal{M}$, $\mathcal{A}$

4.3 Experiment 3: Abbreviation-Aware Coreference Resolution Using LLM-Extracted Dictionaries

Let $A = \{(a_j, \alpha_j)\}$ be abbreviation-definition pairs extracted from the first 750 words using the GPT-4o.

$$R_i = \text{LLM}_{\text{abbr}}(A; p_i)$$

- **Prompt:** "Here is a list of abbreviations A.

the model is requested to extract all the coreferencing categories in separate runs, then, rewrite paragraph p_i by expanding ambiguous abbreviations and resolving references."

- **Purpose:** Leverage explicit abbreviation knowledge to aid disambiguation of biological mentions.

Biomedical texts frequently employ abbreviations for complex names, which can cause substantial ambiguity in coreference resolution. In this experiment, we assess whether providing LLMs with a structured abbreviation dictionary improves coreference resolution compared to unstructured context, such as the reference paragraphs used in Experiment 2. To build this dictionary, we parse the first 750 wordsof each document using Stanza and extract abbreviation-definition pairs (e.g., APP = "amyloid precursor protein") using the GPT-4o interface. These pairs are then validated against the CRAFT corpus. The resulting Abbreviation List is passed as auxiliary input during prompting.

4.4 Experiment 4: Entity-Aware Coreference Resolution Using LLM-Extracted Dictionaries

Let $E = \{e_k\}$ be key biomedical entities extracted from the first 750 words using GPT4o.

$$R_i = \text{LLM}_{\text{entity}}(E; p_i)$$

- **Prompt:** "Here is a list of detected biomedical entities E. Extract all the coreferencing mentions, then, rewrite paragraph pi by expanding ambiguous abbreviations and resolving references"

- **Purpose:** Provide broader semantic grounding than abbreviations alone, to evaluate whether entity awareness supports coherent coreference resolution.

In this experiment, we examine whether incorporating explicit biomedical entity information into the prompting process can enhance the performance of large language models on coreference resolution. Instead of only relying on implicit context or abbreviation mappings, we provide the LLM with a curated Entity List; a list of biomedical terms extracted using GPT-4o interface from the first 750 words of each document and validated against the CRAFT corpus to ensure correctness.

Table 2: LLaMA models performance metrics for LOCAL and REF_CTX tasks

Size	Task	P	R	F1
70B	LOCAL	0.800	0.458	0.583
	REF_CTX	0.805	0.390	0.525
17B	LOCAL	0.825	0.613	0.704
	REF_CTX	0.850	**0.573**	**0.685**
8B	LOCAL	**0.874**	**0.723**	**0.791**
	REF_CTX	**0.906**	0.539	0.676

This entity list serves as a form of semantic grounding. For each paragraph in the input article, the LLM is prompted with both the paragraph and the corresponding entity list. The model is then asked to resolve any ambiguous mentions by aligning them with the most probable entry in the entity list and rewriting the paragraph accordingly.

5 Results Analysis

The span-based baseline SpanBERT-large achieves an F1 of only 0.1322, highlighting the difficulty of biomedical coreference resolution for traditional models constrained by limited context windows, domain-specific terminology, and a mismatch between general-domain fine-tuning and specialized biomedical discourse.

To ensure accurate evaluation, we removed 9,335 gold-standard annotations from the selected CRAFT articles that lacked relation entries as the ones that don't connect to original source cannot be validates for coreferencing resolution. We then matched predicted resolutions against the remaining ∼83,608 annotated spans using case-insensitive partial character overlap ($\geq$2 characters). Predictions were extracted from structured JSON when available, or via a fallback regex parser. Precision, recall, and F1 were computed at the mention level, with unmatched predictions treated as false positives and missed gold spans as false negatives.

Our results reveal consistent patterns in how generative LLMs perform on biomedical coreference resolution:

Model Scale vs. Effectiveness.
Surprisingly, the 8B and 17B LLaMA models outperform the 70B variant across all experiments, indicating that scale alone does not drive coreference performance in domain-specific, fine-grained

Table 3: LLaMA models performance metrics for `abb_dictionary` and `entity_dictionary` tasks

Size	Task	P	R	F1
70B	ABBR	0.844	0.395	0.538
	ENTITY	0.826	0.379	0.519
17B	ABBR	**0.919**	0.400	0.558
	ENTITY	**0.891**	**0.633**	**0.740**
8B	ABBR	0.868	**0.653**	**0.745**
	ENTITY	0.882	0.551	0.678

Table 4: LLaMA models performance metrics for distance-aware local coreference resolution

Size	P	R	F1
70B	0.794	0.430	0.557
17B	0.841	0.554	0.668
8B	**0.892**	**0.601**	**0.718**

reasoning tasks. A likely explanation is that smaller models generalize more conservatively and commit fewer overconfident errors, whereas larger models, despite stronger generative capacity, appear more susceptible to prompt misalignment and semantic overreach.

To further probe local-only performance, we ran an additional variant of Experiment 1 in which paragraphs were selected not by a sequential 200-word window but by the most frequent reference distance observed in CRAFT (500 words), testing whether LLMs benefit from input chunks that align with natural coreference distances. Results appear in Table 4.

Impact of Coreference Distance.
The distance-aware local variant yields only marginal precision gains and a noticeable drop in recall and F1 for LLaMA 17B and 8B, indicating that proximity alone is insufficient for robust coreference; many biomedical entities require contextual cues beyond sentence-local information. LLaMA 8B still achieves the highest F1, confirming its strong span-level sensitivity, yet the degradation with larger context corroborates what we observe in Experiment 2 (`REF_CTX`).

Reference Context Has Mixed Effects.
In Experiment 2, which incorporates a fixed reference paragraph to aid disambiguation, recall often drops compared to the purely local setup (Experiment 1), particularly for LLaMA 8B and 70B. This suggests that, without fine-tuning or explicit multi-span integration mechanisms, LLMs may not reliably incorporate reference paragraphs into their reasoning. Prior work has observed that transformer-based models tend to prioritize recent tokens or local context unless explicitly guided otherwise, which could explain the degraded performance here.

Structured Dictionaries Improve Recall.
The abbreviation-aware settings (Experiments 3) yield measurable recall and F1 gains, most visibly for the 8B model, which reaches recall of 0.653 and 0.745 in `ABBR`. When supplied with structured input; abbreviation definitions, even smaller models like the 8B one can more reliably identify correct antecedents. These findings align with prior evidence that structured, grounded prompting improves information extraction, particularly when domain-specific context is required.

Overall, these results highlight both the promise and the current limitations of generative LLMs for biomedical coreference: while auxiliary signals such as abbreviation and entity dictionaries meaningfully boost performance, LLMs still struggle to integrate multi-paragraph context and resolve less explicit coreferences, underscoring the need for domain-specific adaptation or hybrid approaches that combine generative models with symbolic or retrieval-based components.

Coreference Type Sensitivity and Model Behavior.

To better understand how LLMs handle different forms of coreference, we evaluated each model across four categories; pronouns, indefinite noun phrases, abbreviations, and definite noun phrases, under the `LOCAL`, `REF_CTX`, `ABBR`, and `ENTITY` setups. Because CRAFT does not label coreference types, we built a post-processing pipeline that classifies predicted and gold mentions via lexical heuristics: pronouns, indefinite expressions, and abbreviations are identified by exact matches against dictionaries we manually compiled (available on our GitHub), and the remaining mentions are treated as definite noun phrases. Per-type precision, recall, and F1 are then computed from the model's original prediction labels.

As shown in Figure 3, pronoun coreference consistently yields the highest F1 across all LLaMA

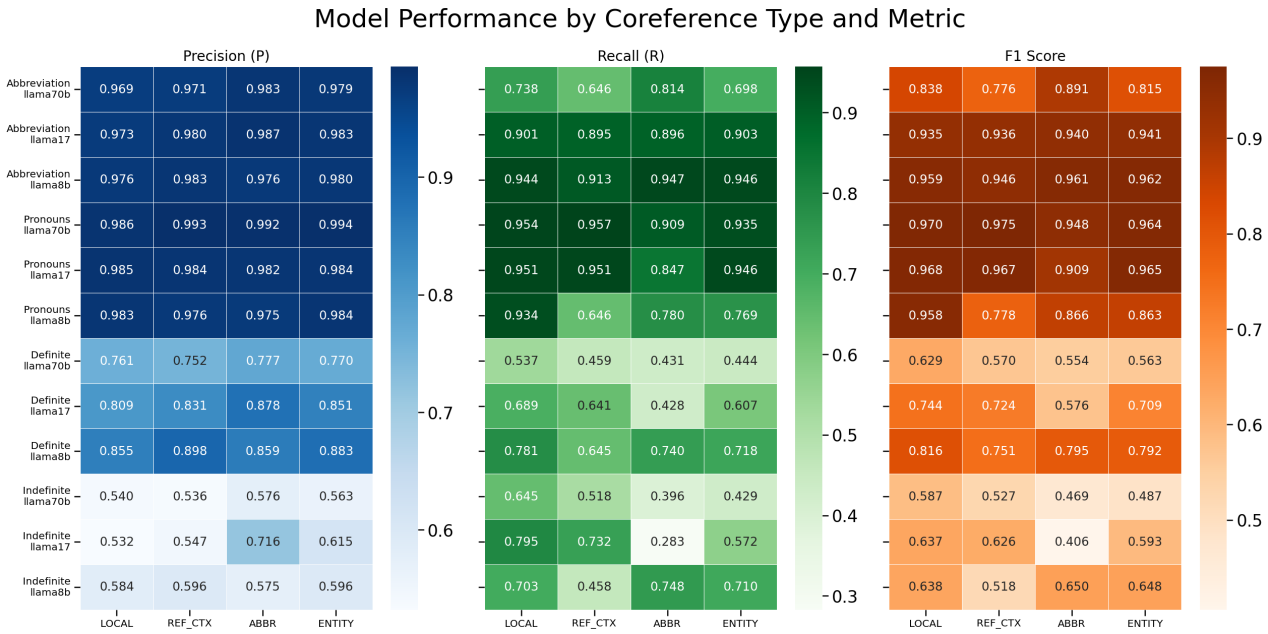


Figure 3: Heatmap of precision, recall, and F1 scores for LLaMA models (70B, 17B, 8B) across four experimental setups (LOCAL, REF CTX, ABBR, ENTITY) and coreference categories (pronouns, indefinite NPs, abbreviations, definite NPs).

models, peaking at 0.975 for LLaMA 70B under REF_CTX, likely because pronouns are heavily represented in pretraining corpora and rely on short-range syntactic cues; complementary evidence from Figure 4 in Appendix A shows pronouns are also resolved in high absolute counts, especially by LLaMA 17B under minimal context.

Abbreviation coreference is similarly strong, particularly under ABBR and ENTITY: injecting abbreviation dictionaries produces noticeable gains in both F1 (e.g., 0.961 for LLaMA 8B in ABBR) and resolved-mention counts, confirming that domain-specific cues substantially aid biomedical abbreviation understanding.

Definite noun phrase resolution reaches moderate F1, with LLaMA 8B highest in LOCAL (F1: 0.816), but both metrics and counts decline slightly under ABBR and ENTITY, likely due to contextual noise from input augmentation, suggesting that local syntactic proximity matters more for definite noun phrases than external cues. Indefinite noun phrases pose the greatest challenge: F1 is lowest across all models and experiments, with LLaMA 17B marginally stronger in REF_CTX and LOCAL. Notably, although LLaMA 70B leads in pronoun F1, it consistently underperforms elsewhere in both metrics and counts, indicating that larger models may be more prone to distraction or overfitting in domain-specific settings.

6 Conclusion

Our study presented a systematic evaluation of generative large language models (LLMs) for coreference resolution in the biomedical domain. We benchmarked three LLaMA models across four prompt-based settings and compared them to a span-based baseline, using the richly annotated CRAFT corpus for evaluation.

Our findings show that LLMs can resolve coreference with reasonable precision, but often suffer from low recall without explicit cues. Notably, smaller models like LLaMA 8B frequently outperformed larger ones, especially when provided with structured input such as abbreviation. In contrast, injecting additional free-text context offered limited benefit and sometimes degraded performance.

Overall, these results highlight the nuanced relationship between model size, coreference category, and the design of contextual input. They emphasize that targeted domain-specific augmentation, such as structured dictionaries, can have a greater impact on performance than model scale alone. Notably, smaller models can match or even exceed the performance of larger ones when paired with carefully designed prompts. Future directions should explore fine-tuning strategies, integration of external biomedical knowledge, and hybrid generative extractive systems to further enhance recall and robustness.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. https://arxiv.org/abs/1912.01140. Proceedings of the 12th Language Resources and Evaluation Conference (LREC), pages 44–54, Marseille, France.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

K. Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter. 2017a. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. https://doi.org/10.1186/s12859-017-1775-9. BMC Bioinformatics 18(1):372.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner Jr, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017b. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):372.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. *arXiv preprint arXiv:2109.04127*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1971–1982.

Yujian Gan, Juntao Yu, and Massimo Poesio. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 1645–1665. European Language Resources Association (ELRA).

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Yufei Li, Xiangyu Zhou, Jie Ma, Xiaoyong Ma, Pengzhen Cheng, Tieliang Gong, and Chen Li. 2022. Distinguished representation of identical mentions in bio-entity coreference resolution. *BMC Medical Informatics and Decision Making*, 22(1):116.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun Wang, Jianwei Yin, and Xuhong Zhang. 2024. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. *arXiv preprint arXiv:2410.01671*.

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. *arXiv preprint arXiv:2109.12424*.

Kawshik Manikantan, Makarand Tapaswi, Vineet Gandhi, and Shubham Toshniwal. 2024. IdentifyMe: A Challenging Long-Context Mention Resolution Benchmark for LLMs. https://arxiv.org/abs/2411.07466. ArXiv preprint arXiv:2411.07466.

Meta AI. 2024a. Meta Llama 3.1 8B Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Released July 23, 2024, under the Llama 3.1 Community License.

Meta AI. 2024b. Meta Llama 3.3 70B Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Released July 23, 2024, under the Llama 3.1 Community License.

Meta AI. 2025. Meta Llama 4 Scout 17B-16E Instruct. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Released April 5, 2025, under the Llama 4 Community License.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111.

Brendan O'Connor and Michael Heilman. 2013. Arkref: A rule-based coreference resolution system. *arXiv preprint arXiv:1310.1975*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. https://aclanthology.org/W13-3516/. Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL), pages 143–152.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Nourah M. Salem and 1 others. 2025. BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs. https://arxiv.org/abs/2510.25087. ArXiv preprint arXiv:2510.25087.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Sam Joshua Wiseman, Alexander Sasha Rush, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6953–6963.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, and 1 others. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

7 Appendix A

7.1 Additional Experiment Results

Figure 4 reports the absolute number of coreference mentions extracted by each LLaMA model (70B, 17B, 8B) under the four experimental setups (LOCAL, REF_CTX, ABBR, ENTITY), broken out by coreference category: pronouns, indefinite noun phrases, abbreviations, and definite noun phrases. The panel-wise y-axis ranges differ substantially across categories, reflecting the relative frequency with which each coreference type occurs in the CRAFT articles; definite noun phrases and abbreviations are by far the most numerous classes, while pronouns and indefinite noun phrases appear at an order of magnitude smaller scale.

Two patterns stand out. First, pronoun and definite-noun-phrase extraction are dominated by the larger models under minimal context, with LLaMA 70B and LLaMA 17B producing the highest counts in the LOCAL and REF_CTX conditions and then tapering off under the dictionary-augmented prompts, consistent with their high pronoun F1 scores in Figure 3 and the short-range syntactic nature of these mentions. Second, abbreviation extraction shows the opposite trend: LLaMA 8B produces the largest number of resolved abbreviations, and its count grows sharply under the ABBR and ENTITY prompts, mirroring the 0.961 F1 it achieves on abbreviations in the ABBR setup. Indefinite noun phrases follow a middle pattern, with LLaMA 17B leading in the lower-context conditions (LOCAL, REF_CTX) and all three models converging once structured dictionaries are supplied. Together these counts reinforce the quantitative findings from Tables 2 and 3: structured domain cues disproportionately benefit the smaller 8B model, particularly on the abbreviation-heavy categories where biomedical grounding matters most.

7.2 Computational Resouces

All open-weight LLM experiments were conducted on a high-performance Google Cloud VM instance of type `a2-highgpu-8g`, equipped with 96 vCPUs, 680 GB of RAM, and 8 NVIDIA A100 GPUs (40 GB each). The environment was provisioned with the c0-deeplearning-common-cu118-v20241118-debian-11-py310 boot disk image, ensuring compatibility with CUDA 11.8 and PyTorch 2.x frameworks.

8 Appendix B

8.1 Prompt Template

To guide the language model's behavior consistently across experiments, we employ a structured system prompt for each coreference type, and this prompt instructs the model to identify and resolve only a targeted subset of coreference expressions. In this example, the focus is on definite noun phrase coreferences within a paragraph, while explicitly excluding pronouns, indefinite expressions, and abbreviations.

The same system-prompt skeleton is reused across all four experimental setups, with only the

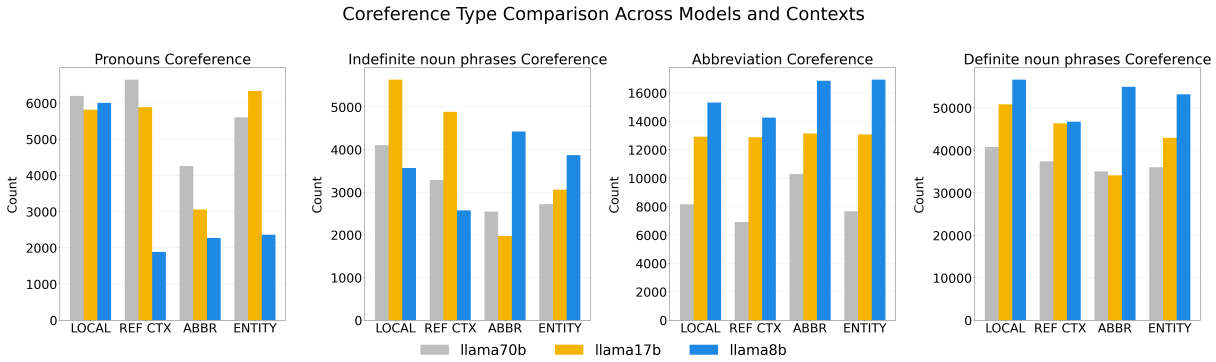


Figure 4: Extracted coreference type counts by model and context.

auxiliary-context block changing: LOCAL uses the skeleton verbatim, REF_CTX prepends the first paragraph of the article as a fixed reference block, ABBR injects the extracted abbreviation–definition pairs $\mathcal{A}$, and ENTITY injects the extracted biomedical entity list $\mathcal{E}$. The abbreviation and entity dictionaries used in Experiments 3 and 4 are themselves produced by a separate GPT-4o extraction prompt applied to the first 750 words of each document and validated against the CRAFT gold annotations. The four coreference categories (pronouns, definite noun phrases, indefinite noun phrases, and abbreviations) are resolved in independent inference runs; only the "ONLY X" clause and the accompanying skip list change between runs, while the JSON output schema remains identical. All LLaMA inferences are run with temperature 0 and top-$p = 1$ to maximize determinism, with `max_new_tokens` set to accommodate the full rewritten paragraph plus its JSON envelope. Model responses are first parsed as strict JSON; when parsing fails, a fallback regex extractor recovers the `Extracted_Expressions`, `Resolutions`, and `Rewritten_Paragraph` fields and the example is retained only if all three are present. Malformed outputs that cannot be recovered are discarded and counted as empty predictions for evaluation.

System Prompt

You are a scientific language model with expert-level understanding of coreference resolution.
Your task is to extract and resolve ONLY **definite noun phrase coreferences** (e.g., "the gene", "these proteins", "such results") within the paragraph.
Skip the following:

- Pronouns (e.g., "it", "they")
- Indefinite noun phrases (e.g., "a result", "some proteins")
- Abbreviations (e.g., "IOP")

Follow these steps:

1. Extract coreference expressions that appear *verbatim* in the paragraph. **Do NOT invent or rephrase them**.
2. For each expression, resolve it to its correct antecedent using context from the same paragraph.
3. Rewrite the paragraph by substituting each extracted expression with its resolved referent.

DO NOT paraphrase, summarize, add, remove, or reorder any content. **Preserve the original wording and sentence structure, except for the substitutions.**

Expected JSON Output Schema:

```
{
        "Extracted_Expressions": [
        "[expression_1]",
        "[expression_2]"
        ],
        "Resolutions": {
                "[expression_1]": "[detailed
    explanation describing the antecedent]",
                "[expression_2]": "[detailed
    explanation describing the antecedent]"
        },
        "Rewritten_Paragraph": "[the rewritten
    paragraph, identical except for
    substitutions]"
}
```

Example:

> *Input:* "These results were unexpected. They indicate a new trend."
> *Rewritten:* "The XYZ experiment results were unexpected. The results indicate a new trend."

Example Output:

```
{
        "Coreference_Resolution": {
                "Extracted_Expressions": [
                "IOP",
                "IOPs",
                "They"
                ],
                "Resolutions": {
                        "IOP": "intraocular
    pressure",
                        "IOPs": "intraocular
    pressures",
                        "They": "Genetically
    distinct mouse strains"
                },
                "Rewritten_Paragraph": "
    Intraocular pressure in genetically distinct
     mice: an update and strain survey..."
           }
}
```

A Multi-Agent Open-Source LLM for Structured Cancer Registry Information Extraction from Pathology and Medical Reports

Riham Jeeballah[1], **Adhari Al Zaabi**[2], **Habiba El Keraby**[1], **and Abdulrahman AAl Abdulsalam**[1,*]

Corresponding author: a.aalabdulsalam@squ.edu.om

[1]Sultan Qaboos University, Department of Computer Science, Oman

[2]Sultan Qaboos University, Department of Clinical and Human Anatomy, Oman

Abstract

Extracting structured cancer registry information from pathology and medical reports is challenging due to heterogeneous reporting styles and implicit clinical reasoning. We propose a modular multi-agent framework that decomposes registry abstraction into semantic chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation stages.

The dataset includes 818 annotated cancer cases from Sultan Qaboos University Hospital. Evaluation in this study focuses on breast (n=454) and colorectal (n=174) reports across grade, morphology, TNM staging, and laterality extraction tasks. The framework is compared against prompt-based LLaMA 3.3 baselines using accuracy and weighted/macro F1-score metrics.

The proposed framework improved performance in context-dependent tasks, particularly grade extraction, where weighted F1-score increased from 0.71 to 0.78 for breast cancer and from 0.56 to 0.67 for colorectal cancer. Improvements were also observed for colorectal laterality extraction. For other extraction tasks, particularly highly structured tasks such as TNM staging and morphology extraction, the multi-agent framework achieved performance comparable to direct prompting.

Although the baseline achieved slightly higher average weighted F1-scores overall, the proposed framework provides improved modularity, traceability, and pipeline-level interpretability through explicit intermediate reasoning stages, supporting error analysis and future clinician-guided refinement.

1 Introduction

Cancer registries are essential for epidemiological surveillance, clinical research, and healthcare planning. However, key registry variables such as grade, morphology, TNM staging, and laterality are often embedded in free-text pathology reports, making abstraction labor-intensive, difficult to scale, and prone to inter-annotator variability.

Extracting structured information from clinical narratives remains challenging due to heterogeneous documentation styles, implicit terminology, and contextual ambiguity. While transformer models and large language models (LLMs) improved clinical text understanding and structured output generation, single-pass prompting approaches remain vulnerable to hallucination, inconsistency, and weak grounding when extracting multiple interdependent variables.

Recent advances in agentic LLM systems offer an alternative by decomposing complex tasks into modular reasoning stages. The ReAct paradigm demonstrated that combining reasoning and action improves performance and interpretability (Yao et al., 2022). Building on this idea, biomedical agentic systems incorporated planning, orchestration, and tool use (Abbasian et al., 2024; Li et al., 2025), while surveys and benchmark studies characterized these paradigms in biomedicine (Lin et al., 2024; Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Gorenshtein et al., 2025). In oncology, agentic frameworks have been explored for decision support, diagnostic reasoning, and treatment planning (Ferber et al., 2025; Chen et al., 2025; Çağatay Umut Öğdü et al., 2025; Kuerbanjiang et al., 2025; Inoue et al., 2025; Zhao et al., 2025a). In parallel, LLM-based approaches have also been investigated for structured extraction from pathology and oncology reports (Chow et al., 2025; Gupta et al., 2025; Hart and Bergamaschi, 2026).

Despite this progress, existing approaches often focus on isolated tasks, single cancer types, or direct prompting strategies with limited visibility into intermediate reasoning stages and failure sources.

In this work, we propose an open-source multi-agent framework for structured cancer registry extraction from pathology and medical reports,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 531–551
July 3-4, 2026 ©2026 Association for Computational Linguistics

focusing on breast and colorectal cancer. Implemented as a sequential LangGraph[1] workflow, the framework performs semantic chunking, field-conditioned retrieval, field-specific extraction, validation, evaluation, and aggregation. Unlike direct prompting, the system stores intermediate outputs across stages, enabling pipeline-level traceability and allowing failures to be linked to specific stages such as retrieval, extraction, normalization, or aggregation. This design supports structured auditing and future clinician-guided refinement for clinical information extraction.

2 Related Work

2.1 Clinical Information Extraction from Pathology Reports

Clinical information extraction evolved from rule-based systems to transformer-based and LLM-based approaches. Although transformer models improved generalization across tasks such as named entity recognition and relation extraction, pathology reports remain challenging due to dense terminology, implicit structure, and contextual ambiguity.

Recent studies demonstrated that LLMs can perform structured extraction from clinical and pathology documents without extensive task-specific training. Chow et al. (2025) demonstrated strong performance of open-weight LLMs for schema-aligned pathology abstraction. Hart and Bergamaschi (2026) proposed an agent-based framework for breast cancer pathology extraction, while Gupta et al. (2025) introduced a hierarchical agentic system for large-scale oncology data extraction. However, these approaches mainly focus on single-task extraction, institution-specific workflows, or direct prompting strategies with limited pipeline-level interpretability.

2.2 LLM Agents in Biomedicine and Oncology

Agentic AI expands the capabilities of LLMs by enabling iterative reasoning, external tool interaction, memory utilization, and coordinated multi-step problem solving (Yao et al., 2022; Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Zhao et al., 2025b; Abdollahi et al., 2025). Foundational work such as ReAct demonstrated the advantages of integrating reasoning with action execution (Yao et al., 2022), while subsequent studies proposed biomedical agentic frameworks incorporating orchestration and tool-assisted workflows (Abbasian et al., 2024; Lin et al., 2024; Li et al., 2025; Inoue et al., 2025). Recent surveys further emphasized the growing applications of agentic AI in clinical decision-making, research automation, and biomedical knowledge synthesis (Xu and Sankar, 2025; Yang et al., 2025; Radi et al., 2025; Zhao et al., 2025b; Abdollahi et al., 2025).

In oncology, agentic frameworks have been explored for decision support, diagnostic reasoning, treatment planning (Ferber et al., 2025; Chen et al., 2025; Kuerbanjiang et al., 2025; Liu et al., 2026), and biomedical research applications such as drug discovery and biomarker reasoning (Inoue et al., 2025; Zuo et al., 2025; Jin et al., 2025; Huang et al., 2025; Bazgir et al., 2025). However, most work focuses on decision support or research workflows rather than structured cancer registry extraction from heterogeneous pathology documentation.

2.3 Research Gap

Despite progress in clinical information extraction and agentic AI, several gaps remain at their intersection. Existing approaches often rely on direct prompting strategies that process entire reports in a single inference step. While effective for explicit patterns such as TNM expressions, these approaches provide limited visibility into intermediate reasoning stages and extraction failures. In cancer registry abstraction, clinically relevant evidence may be distributed across multiple report sections, expressed implicitly, or embedded within noisy longitudinal documentation.

To address these limitations, we propose a modular multi-agent framework implemented as a sequential LangGraph workflow that performs semantic chunking, field-conditioned retrieval, field-specific extraction, validation, evaluation, and aggregation. The framework additionally stores intermediate outputs across stages, enabling pipeline-level traceability, structured auditing, and future clinician-guided refinement.

3 Methodology

3.1 Study Overview

We propose a modular multi-agent framework for structured cancer registry extraction from pathology and medical reports, targeting clinically relevant variables including grade, morphology, TNM staging, and laterality. Unlike single-pass LLM

[1] https://github.com/langchain-ai/langgraph

prompting, the framework decomposes extraction into sequential specialized stages implemented as a directed LangGraph workflow. Agents operate over a shared state and progressively refine intermediate outputs, enabling modular execution, structured auditing, and pipeline-level traceability.

3.2 Dataset

The dataset was collected from Sultan Qaboos University Hospital and includes 818 annotated cancer cases. Clinical reports were independently reviewed by trained pathologists, with disagreements adjudicated by a senior pathologist to establish the final gold standard annotations. Inter-annotator agreement analysis across registry variables and cancer types is provided in Appendix B.

Although the overall dataset contains multiple cancer types, the current study focuses on breast (n=454) and colorectal (n=174) reports for evaluation across four registry variables: grade, morphology, TNM staging, and laterality.

3.3 Multi-Agent Architecture

The proposed system consists of sequential specialized agents, as illustrated in Figure 1. Additional implementation details and source code availability are provided in Code Availabilty section.

- **Chunking Agent:** Segments reports into semantically coherent chunks and assigns field-relevance labels.
- **Retriever Agent:** Selects field-specific evidence relevant to grade, morphology, TNM staging, and laterality extraction. Retrieval is used to reduce irrelevant context and improve extraction focus for context-dependent variables.
- **Field-Specific Extraction Agents:** Dedicated agents independently extract grade, morphology, T, N, M, and laterality using task-specific prompts and structured output constraints.
- **Reviewer Agent:** Validates outputs, enforces schema consistency, and repairs malformed predictions.
- **Evaluator Agent:** Computes evaluation metrics when reference labels are available, or estimates confidence using evidence alignment and internal consistency.
- **Aggregation Agent:** Consolidates validated outputs into a unified structured record containing predictions, normalized labels, confidence scores, and supporting evidence.

All prompts are provided in Appendix D. Prompt design was iteratively refined through experimental evaluation and error inspection to improve schema consistency and extraction reliability across fields.

3.4 Workflow Orchestration

The pipeline is implemented using LangGraph as a directed execution graph operating over a shared state containing reports, retrieved evidence, intermediate outputs, predictions, and confidence scores. This design enables modular execution, transparent information flow, and stage-level error tracing.

3.5 Model Communication Protocol (MCP)

We introduce a lightweight Model Communication Protocol (MCP) that standardizes communication between agents through unified schemas for inputs, outputs, intermediate states, and metadata. MCP supports consistent state updates, logging, and reproducible execution across the pipeline.

3.6 Persistent Audit Database

A SQLite-based persistence module stores reports, intermediate outputs, predictions, and confidence values across runs, supporting auditing, traceability, and future refinement.

3.7 Implementation Details

The framework is implemented in Python using LangGraph. LLaMA 3.3 (70B) is deployed locally through Ollama (v0.4.7) to preserve data privacy. The system runs on Ubuntu 22.04 LTS with Python 3.10, Streamlit for visualization, and Pandas for data processing. Experiments were conducted using two NVIDIA RTX 6000 Ada GPUs (48 GB each) with CUDA 13.0.

3.8 Baseline Method

As a baseline, we implement direct prompt-based extraction using LLaMA 3.3 without task decomposition, retrieval, or intermediate reasoning stages. Multiple prompting strategies were evaluated, including simple/prefix prompting (direct extraction instructions), cloze prompting (fill-in-the-blank style extraction), anticipatory prompting (guiding the model with preparatory contextual instructions), chain-of-thought prompting (step-wise reasoning

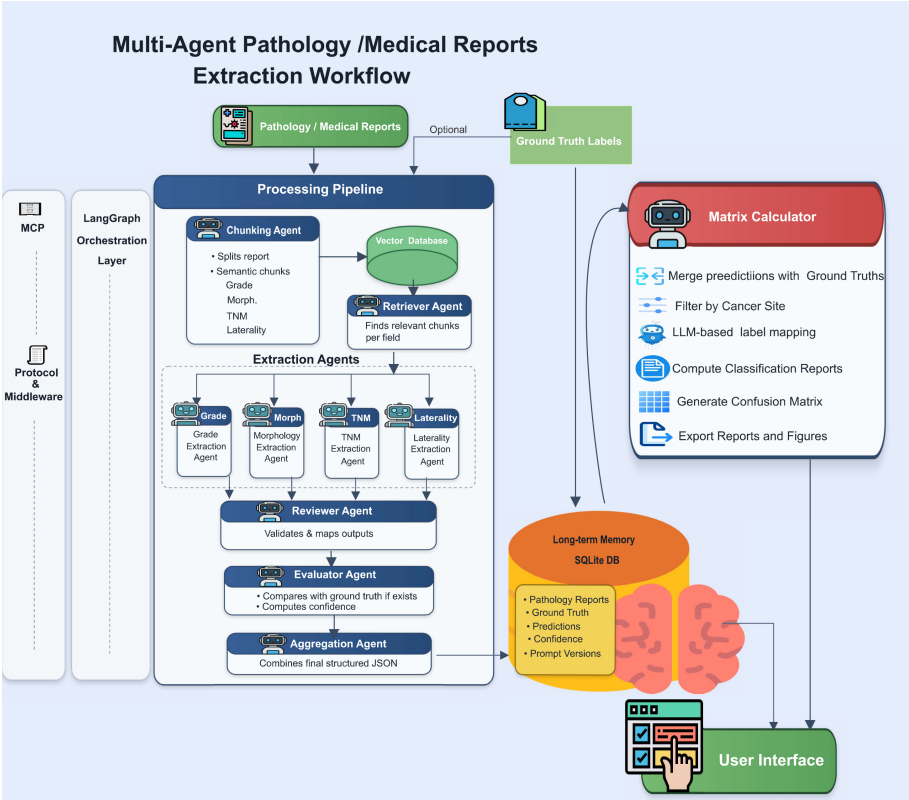


Figure 1: Overview of the proposed multi-agent framework. The pipeline consists of sequential agents for chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation coordinated through a shared state.

before prediction), and heuristic prompting (rule-guided extraction instructions). For each extraction task, the best-performing baseline prompting strategy was selected for comparison against the proposed multi-agent framework. Detailed baseline prompts are provided in Appendix C.

Compared to direct prompting, the proposed multi-agent framework introduces additional computational overhead due to sequential agent execution, retrieval stages, and intermediate validation steps. However, the current study primarily focuses on extraction quality, interpretability, and traceability rather than runtime optimization.

3.9 Evaluation Metrics

Performance is evaluated using accuracy, macro-averaged, and weighted precision, recall, and F1-score. Macro metrics assess performance equally across classes, while weighted metrics account for class imbalance and reflect real-world clinical distributions.

Predictions are evaluated using exact-label matching against ground-truth registry annotations.

4 Results and Discussion

This section compares the proposed multi-agent framework with prompt-based LLaMA 3.3 baselines for breast and colorectal cancer registry extraction across grade, morphology, TNM staging, and laterality tasks. Because several fields are imbalanced, weighted F1-score is used as the primary metric, while macro metrics evaluate class-balanced performance.

Overall, results reveal task-dependent behavior. The multi-agent framework improves context-dependent tasks such as grade extraction and colorectal laterality, while achieving performance comparable to direct prompting for highly structured fields such as TNM staging and morphology.

4.1 Grade Extraction Performance

Table 1 summarizes grade extraction performance across breast and colorectal cancer sites. Figure 2 presents weighted metrics, while weighted F1-only and macro-level comparisons are provided in Appendix A (Figures 8 and 9).

For breast cancer, the multi-agent framework improved weighted F1-score from 0.71 to 0.78 and accuracy from 0.64 to 0.73. For colorectal cancer,

Table 1: Comparative performance for grade extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Anticipatory	0.64	0.83	0.64	0.71	0.55	0.61	0.51
Breast	Multi-Agent	0.73	0.83	0.73	0.78	0.58	0.54	0.56
Colorectal	Baseline-Anticipatory	0.44	0.85	0.44	0.56	0.40	0.34	0.34
Colorectal	Multi-Agent	0.59	0.80	0.59	0.67	0.34	0.30	0.31

Table 2: Comparative performance for morphology extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Anticipatory	0.75	0.91	0.75	0.82	0.34	0.40	0.33
Breast	Multi-Agent	0.72	0.92	0.72	0.80	0.35	0.54	0.34
Colorectal	Baseline-Simple	0.86	0.87	0.86	0.87	0.31	0.31	0.31
Colorectal	Multi-Agent	0.67	0.99	0.67	0.80	0.33	0.23	0.27

weighted F1-score increased from 0.56 to 0.67 and accuracy from 0.44 to 0.59. These improvements suggest that decomposition and field-specific evidence retrieval are beneficial when grade-related information is context-dependent or distributed across multiple report sections.

However, the slight decrease in colorectal macro F1-score indicates reduced sensitivity for less frequent classes. Overall, grade extraction represents the strongest advantage of the proposed multi-agent framework over direct prompting.

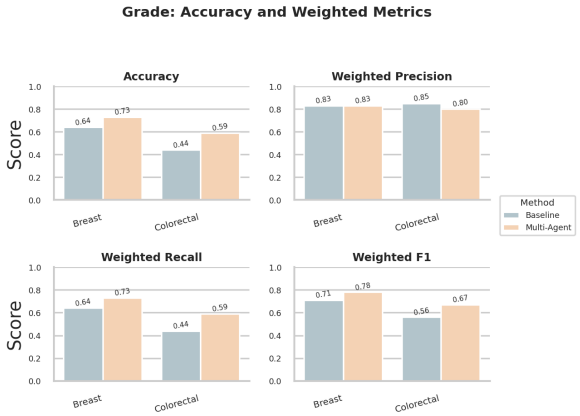


Figure 2: Accuracy, weighted precision, weighted recall, and weighted F1-score comparison between the baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

4.2 Morphology Extraction Performance

Table 2 and Figure 3 summarize morphology extraction performance. Additional weighted F1-only and macro-level comparisons are provided in Figures 10 and 11. For breast cancer, the multi-agent framework remained close to the baseline, with weighted F1-score decreasing slightly from 0.82 to 0.80 while macro F1-score improved marginally from 0.33 to 0.34. In colorectal cancer, the baseline achieved higher weighted F1-score, increasing from 0.80 to 0.87 (+0.07) compared to the multi-agent framework.

These findings suggest that morphology extraction often depends on explicit histologic terminology that can already be captured effectively through direct prompting. Additional retrieval and normalization stages may occasionally introduce information loss or normalization inconsistencies, particularly for less frequent morphology classes.

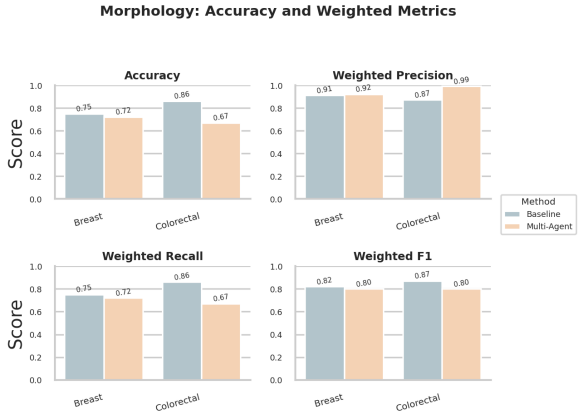


Figure 3: Accuracy, weighted precision, recall, and weighted F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

4.3 TNM Extraction Performance

Table 3 and Figure 4 summarize TNM extraction performance. Additional weighted F1-only and macro-level analyses are provided in Figures 12, 13, 14, 15, 16, and 17. Across TNM components, the baseline generally achieved stronger weighted performance, particularly for T and M categories. This is expected because TNM staging often appears in concise and standardized forms that are well suited to direct prompting. In contrast, the multi-agent framework introduces additional chunking, retrieval, and normalization stages that may occasionally reduce useful contextual information.

4.3.1 T Category

For T-stage extraction, the baseline outperformed the multi-agent framework in both cancer sites. In breast cancer, weighted F1-score decreased from 0.75 to 0.66, while in colorectal cancer it decreased slightly from 0.60 to 0.59. The T sub-figure in Figure 4 confirms that the baseline achieves stronger weighted metrics across both sites, particularly in

breast cancer. These findings suggest that T-stage extraction relies heavily on concise and standardized expressions that can already be captured effectively through direct prompting.

4.3.2 N Category

For N-stage extraction, the multi-agent framework achieved performance comparable to the baseline in breast cancer, with weighted F1-scores of 0.62 for both approaches while slightly improving macro F1-score. In colorectal cancer, the baseline achieved higher weighted F1-score, whereas the multi-agent framework showed slightly improved macro-level behavior. The N sub-figure in Figure 4 illustrates this site-dependent trend.

4.3.3 M Category

For M-stage extraction, the baseline consistently outperformed the multi-agent framework across both cancer sites. In breast cancer, weighted F1-score decreased from 0.60 to 0.48, while in colorectal cancer it decreased from 0.66 to 0.48. The M sub-figure in Figure 4 highlights a clear precision–recall trade-off: although the multi-agent framework achieved higher precision in some settings, it suffered from substantially lower recall, suggesting conservative predictions that missed metastasis-related cases.

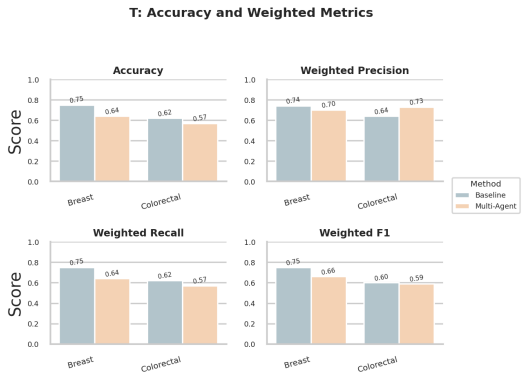
4.4 Laterality Extraction Performance

Table 4 and Figure 5 summarize laterality extraction performance. Weighted F1-only and macro-level comparisons are shown in Figures 18 and 19.

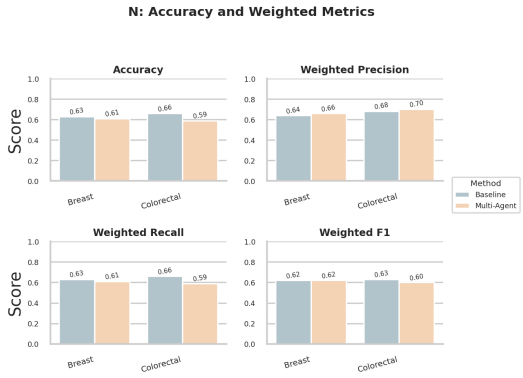
Laterality extraction showed site-dependent behavior. In breast cancer, both approaches achieved high performance, with weighted F1-score of 0.92 for the baseline and 0.91 for the multi-agent framework. Figure 5 confirms that differences across weighted metrics are minimal in this setting.

In colorectal cancer, the multi-agent framework improved weighted F1-score from 0.45 to 0.50 and macro F1-score from 0.22 to 0.27. The corresponding figure suggests that these gains are associated with improved recall and better handling of less explicit laterality expressions.

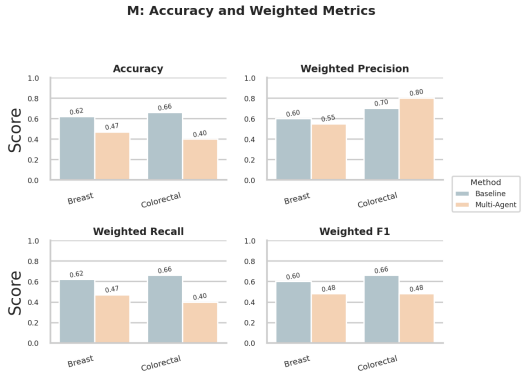
Overall, laterality emerged as one of the more stable extraction tasks, particularly in breast cancer where laterality is often explicitly stated. These findings further suggest that decomposition and field-specific retrieval are beneficial when laterality information is expressed inconsistently or implicitly across report sections.



T-stage



N-stage



M-stage

Figure 4: Accuracy, weighted precision, recall, and weighted F1-score comparison for TNM extraction across breast and colorectal cancer sites.

4.5 Overall Cross-Field Performance Analysis

To assess global system behavior, we analyze weighted F1-scores across all extraction tasks (grade, morphology, TNM, and laterality) using radar plots for each cancer site.

Figures 6 and 7 show that the multi-agent framework redistributes performance across tasks rather than uniformly improving all of them. In breast cancer, the largest gain is observed in grade extraction, while morphology and laterality remain competitive and TNM categories, particularly T

Table 3: Comparative performance for TNM extraction in breast and colorectal cancer.

Cat.	Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
T	Breast	Baseline-Cloze	0.75	0.74	0.75	0.75	0.65	0.60	0.62
T	Breast	Multi-Agent	0.64	0.70	0.64	0.66	0.51	0.47	0.47
T	Colorectal	Baseline-Cloze	0.62	0.64	0.62	0.60	0.61	0.48	0.51
T	Colorectal	Multi-Agent	0.57	0.73	0.57	0.59	0.55	0.46	0.47
N	Breast	Baseline-Cloze	0.63	0.64	0.63	0.62	0.46	0.44	0.43
N	Breast	Multi-Agent	0.61	0.66	0.61	0.62	0.47	0.46	0.45
N	Colorectal	Baseline-Anticipatory	0.66	0.68	0.66	0.63	0.49	0.53	0.47
N	Colorectal	Multi-Agent	0.59	0.70	0.59	0.60	0.58	0.49	0.50
M	Breast	Baseline-Cloze	0.62	0.60	0.62	0.60	0.56	0.50	0.51
M	Breast	Multi-Agent	0.47	0.55	0.47	0.48	0.44	0.47	0.40
M	Colorectal	Baseline-Anticipatory	0.66	0.70	0.66	0.66	0.60	0.58	0.57
M	Colorectal	Multi-Agent	0.40	0.80	0.40	0.48	0.64	0.47	0.40

Table 4: Comparative performance for laterality extraction in breast and colorectal cancer.

Site	Method	Acc	P_w	R_w	$F1_w$	P_m	R_m	$F1_m$
Breast	Baseline-Heuristic	0.90	0.94	0.90	0.92	0.58	0.42	0.45
Breast	Multi-Agent	0.89	0.92	0.89	0.91	0.47	0.46	0.46
Colorectal	Baseline-Simple	0.36	0.72	0.36	0.45	0.27	0.48	0.22
Colorectal	Multi-Agent	0.40	0.73	0.40	0.50	0.35	0.47	0.27

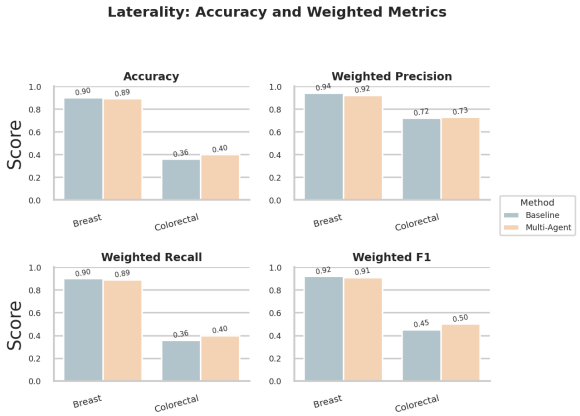


Figure 5: Accuracy, weighted precision, recall, and weighted F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

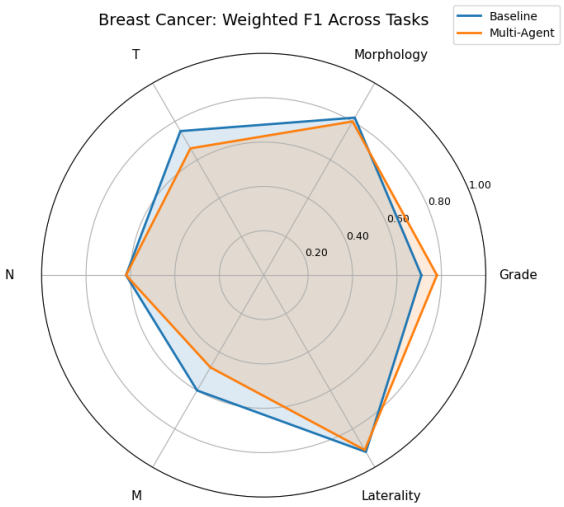


Figure 6: Weighted F1-score comparison across extraction tasks for breast cancer.

Table 5: Average weighted F1-score across all extraction tasks for each cancer site.

Site	Baseline	Multi-Agent
Breast	0.74	0.71
Colorectal	0.63	0.61

and M, favor the baseline. In colorectal cancer, the multi-agent framework improves grade and laterality but underperforms in morphology and M-stage extraction.

Table 5 reports the average weighted F1-score across tasks. The baseline achieves slightly higher overall averages in both breast (0.74 vs. 0.71) and colorectal (0.63 vs. 0.61). However, these aggregate results mask important task-dependent differences and varying levels of clinical complexity.

The baseline performs better in highly structured tasks such as TNM staging and morphology extraction, where clinically relevant information often appears in concise and standardized forms. In contrast, the multi-agent framework demonstrates clear advantages in more context-dependent tasks such as grade extraction and colorectal laterality, where evidence may be distributed across sections or expressed implicitly.

Pipeline-level inspection suggests that performance degradation primarily originates from retrieval and normalization stages rather than extraction itself. For structured fields, chunking and retrieval may occasionally remove useful context or introduce unnecessary normalization steps. In contrast, decomposition improves extraction when contextual reasoning and evidence localization are required.

A formal ablation study isolating the contribution of chunking, retrieval, and normalization stages was not conducted in the current study, as the primary objective was to evaluate the overall

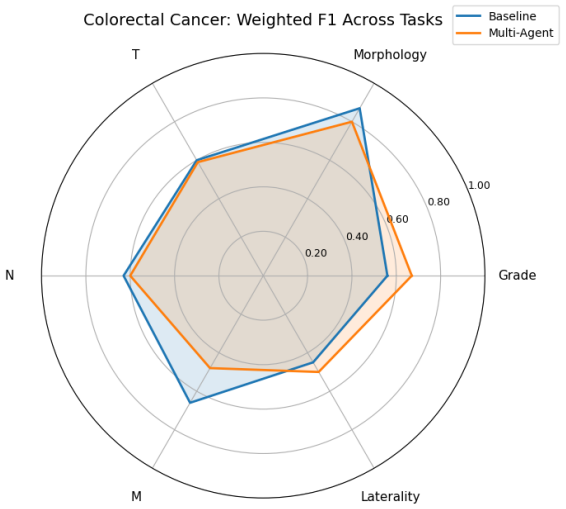


Figure 7: Weighted F1-score comparison across extraction tasks for colorectal cancer.

behavior and clinical applicability of the complete end-to-end framework.

Overall, the proposed framework should not be viewed solely as a score-maximizing alternative, but as a modular and auditable extraction pipeline that supports structured reasoning, traceability, and systematic error analysis.

5 Future Work

Future work will focus on improving retrieval-aware extraction, adaptive chunk selection, and rare-class calibration, particularly for conservative settings such as M-stage extraction where recall degradation was observed. Since performance varies across tasks, future optimization should target field-specific pipeline behavior rather than uniformly increasing model complexity.

Future work will also investigate clinician-guided validation and interactive correction workflows to support targeted refinement of ambiguous predictions. Such feedback may help improve schema normalization, reduce context-related extraction errors, and support iterative prompt and model refinement.

Future evaluation across institutions with different reporting conventions is necessary to better assess generalization and robustness.

Additional directions include evaluating domain-specific biomedical LLMs, comparing alternative prompting strategies more systematically, extending the framework to additional cancer registry fields and cancer sites, and evaluating generalization across institutions with different reporting conventions.

6 Conclusion

In this study, we proposed a modular multi-agent framework for extracting structured cancer registry information from pathology and medical reports. Implemented as a sequential LangGraph workflow, the system decomposes extraction into semantic chunking, retrieval, field-specific extraction, validation, evaluation, and aggregation stages, enabling structured intermediate outputs and pipeline-level traceability.

Experimental results demonstrate that performance is strongly task dependent. The multi-agent framework provides clear advantages in context-dependent tasks such as grade extraction and colorectal laterality, while direct prompting performs better in highly structured tasks such as TNM staging and morphology extraction. These findings highlight an important trade-off between reasoning-driven decomposition and direct pattern-based extraction.

Beyond aggregate performance, the proposed framework provides practical benefits through modular execution, structured auditing, and interpretable intermediate outputs that support error analysis and future refinement. Overall, this work positions multi-agent clinical information extraction as a promising and extensible direction for reliable cancer registry abstraction and future clinically guided NLP systems.

Acknowledgments

This research was funded by Sultan Qaboos University Research Fund under grant number SR-SCI-COMP-22-01.

Code Availability

The implementation of the proposed multi-agent framework, including the LangGraph workflow, extraction agents, evaluation pipeline, and Streamlit interfaces, is publicly available at:
https://github.com/RihamJeeballah/multi-agent-cancer-registry-extraction

References

Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. 2024. Conversational health agents: A personalized llm-powered agent framework.

Ali Abdollahi, Mohammad Amin Rezaei, Xi Wang, Sunan He, Seyed Moein Ayyoubzadeh, Yuxiang Nie, and Hao Chen. 2025. The next pradigm in medical ai: A survey of agentic ai in biomedicine.

Adib Bazgir, Amir Habibdoust Lafmajani, and Yuwen Zhang. 2025. Beyond correlation: Towards causal large language model agents in biomedicine.

Xi Chen, Huahui Yi, Mingke You, Wei Zhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, Qicheng Lao, Weili Fu, Kang Li, and Jian Li. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine*, 8.

Nan-Haw Chow, Han Chang, Hung-Kai Chen, Chen-Yuan Lin, Ying-Lung Liu, Po-Yen Tseng, Li-Ju Shiu, Yen-Wei Chu, Pau-Choo Chung, and Kai-Po Chang. 2025. Comprehensive structured abstraction of pathology reports is now feasible using local large language models.

Dyke Ferber, Omar S.M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, 6:1337–1349.

Alon Gorenshtein, Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Ai agents in clinical medicine: A systematic review.

Shashi Kant Gupta, Arijeet Pramanik, Jerrin John Thomas, Regina Schwind, Lauren Wiener, Avi Raju, Jeremy Kornbluth, Yanshan Wang, Zhaohui Su, and Hrituraj Singh. 2025. Harmon-e: Hierarchical agentic reasoning for multimodal oncology notes to extract structured data.

Steven N. Hart and Teya S. Bergamaschi. 2026. Agent-based large language model system for extracting structured data from breast cancer synoptic reports: a dual-validation study. *JAMIA Open*, 9.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N Carter, Xin Zhou, Matthew Wheeler, Jonathan A Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, and 4 others. 2025. Biomni: A general-purpose biomedical ai agent. *bioRxiv : the preprint server for biology*.

Yoshitaka Inoue, Tianci Song, Xinling Wang, Augustin Luna, and Tianfan Fu. 2025. Drugagent: Multi-agent large language model-based reasoning for drug-target interaction prediction.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. 2025. Stella: Self-evolving llm agent for biomedical research.

Warisijiang Kuerbanjiang, Xinyu Wang, Yiershatijiang Jiamaliding, Musitapa Maimaitiaili, and Yuexiong Yi. 2025. Multidisciplinary large language model agent teams for precision oncology enhance complex gynecologic oncology decision support.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. Agent hospital: A simulacrum of hospital with evolvable medical agents.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z. Li, and Kaicheng Yu. 2024. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, Xianfu Sun, Zhitao Ying, and Guangliang Qiang. 2026. Evomdt: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digital Medicine*.

Abdul Mohaimen Al Radi, Xu Cao, Fanyang Yu, Yuyuan Liu, Fengbei Liu, Chong Wang, Yuanhong Chen, Jintai Chen, Hu Wang, Yanda Meng, Zhenyi Wang, Chen Chen, Mubarak Shah, Tianyu Han, Christos Davatzikos, MacLean P. Nasrallah, and Yu Tian. 2025. Agentic large-language-model systems in medicine: A systematic review and taxonomy.

Xiaoran Xu and Ravi Sankar. 2025. Large language model agents for biomedicine: A comprehensive review of methods, evaluations, challenges, and future directions.

Tiantian Yang, Yihang Xiao, Zhijie Bao, Jianye Hao, and Jiajie Peng. 2025. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models.

Lina Zhao, Jiaxing Bai, Zihao Bian, Qingyue Chen, Yafang Li, Guangbo Li, Min He, Huaiyuan Yao, and Zongjiu Zhang. 2025a. Autonomous multi-modal llm agents for treatment planning in focused ultrasound ablation surgery.

Wenqi Zhao, Shansong Wang, Mojtaba Safari, Mingzhe Hu, and Xiaofeng Yang. 2025b. Medical ai agents: A comprehensive survey of architectures, cognitive modules, and clinical workflows.

Kaiwen Zuo, Zixuan Zhong, Peizhou Huang, Shiyan Tang, Yuyan Chen, and Yirui Jiang. 2025. Healkggen: A hierarchical multi-agent llm framework with knowledge graph enhancement for genetic biomarker-based medical diagnosis.

Çağatay Umut Öğdü, Kübra Arslanoğlu, and Mehmet Karaköse. 2025. An adaptive multi-agent llm-based clinical decision support system integrating biomedical rag and web intelligence. *IEEE Access*, 13:167390–167404.

A Additional Results

This appendix provides supplementary visualizations for weighted F1-only comparisons and macro-level comparisons across extraction tasks.

A.1 Grade Extraction

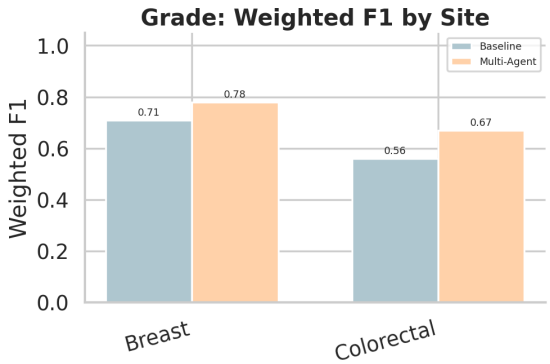


Figure 8: Weighted F1-score comparison between the prompt-based baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

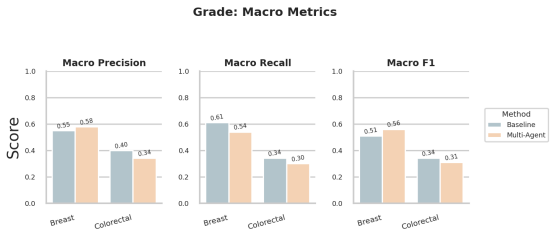


Figure 9: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for grade extraction across breast and colorectal cancer sites.

A.2 Morphology Extraction

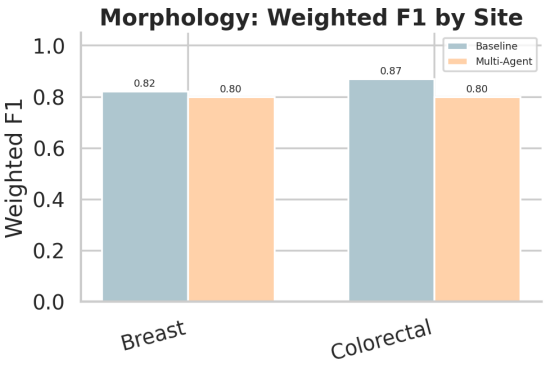


Figure 10: Weighted F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

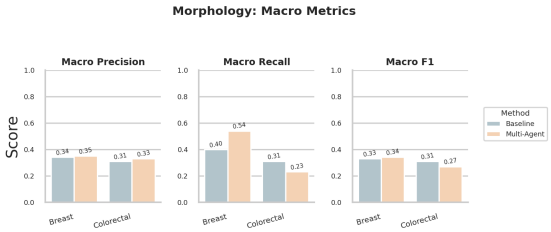


Figure 11: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for morphology extraction across breast and colorectal cancer sites.

A.3 T-Stage Extraction

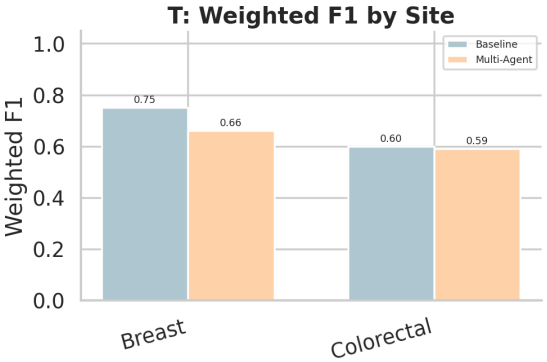


Figure 12: Weighted F1-score comparison between the baseline and the multi-agent framework for T-stage extraction across breast and colorectal cancer sites.



Figure 13: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for T-stage extraction across breast and colorectal cancer sites.

A.4 N-Stage Extraction

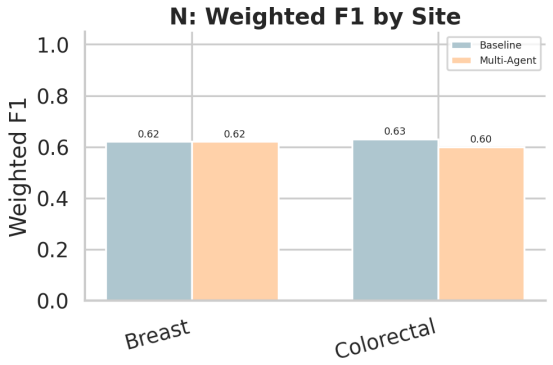


Figure 14: Weighted F1-score comparison between the baseline and the multi-agent framework for N-stage extraction across breast and colorectal cancer sites.

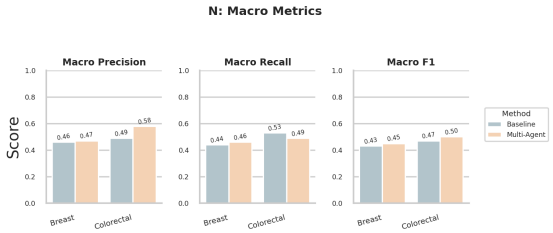


Figure 15: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for N-stage extraction across breast and colorectal cancer sites.

A.5 M-Stage Extraction

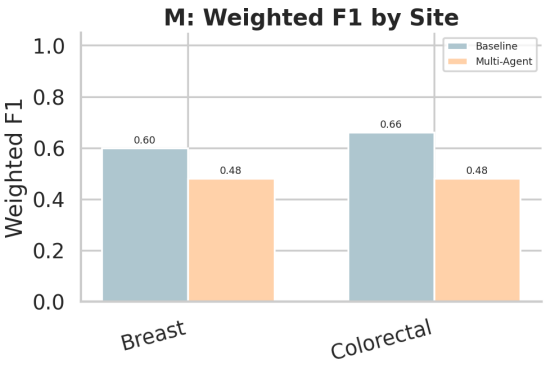


Figure 16: Weighted F1-score comparison between the baseline and the multi-agent framework for M-stage extraction across breast and colorectal cancer sites.

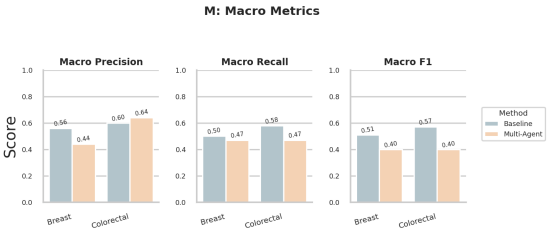


Figure 17: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for M-stage extraction across breast and colorectal cancer sites.

A.6 Laterality Extraction

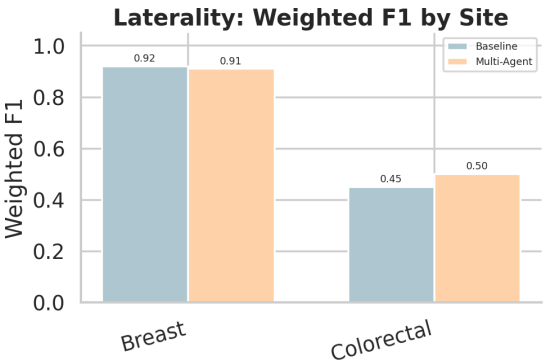


Figure 18: Weighted F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

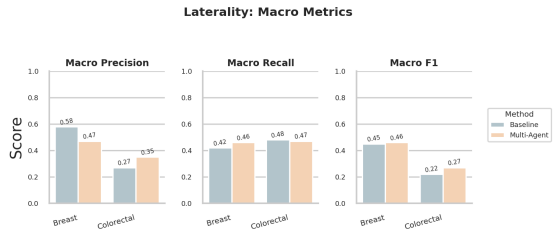


Figure 19: Macro precision, recall, and F1-score comparison between the baseline and the multi-agent framework for laterality extraction across breast and colorectal cancer sites.

B Inter-Annotator Agreement (IAA)

To assess the reliability of the gold-standard annotations, inter-annotator agreement (IAA) was evaluated across all registry variables using Cohen's κ coefficient and percentage agreement. The overall agreement across all fields was $\kappa = 0.917$, with a percentage agreement of 91.99%, indicating high consistency between annotators.

B.1 Overall Agreement Across Registry Fields

Table 6 presents the overall agreement for each registry variable. All fields demonstrate strong agreement, with most achieving "almost perfect" consistency.

Table 6: Overall inter-annotator agreement across registry variables.

Field	Cohen's κ	% Agreement	Interpretation
GRADE	0.968	97.83	Almost perfect
STAGE	0.909	92.05	Almost perfect
MORPHOLOGY	0.981	98.71	Almost perfect
T	0.820	86.80	Almost perfect
N	0.904	92.64	Almost perfect
M	0.812	88.48	Substantial
LATERALITY	0.869	90.33	Almost perfect

B.2 IAA by Cancer Type

Table 7 reports agreement stratified by cancer type. Agreement remains consistently high across most cancer sites, with slight variability observed in more complex fields such as T, N, and M staging, particularly for colorectal and thyroid cases.

C Baseline Prompt Designs

This appendix presents the prompt designs used for the baseline prompt-based LLM extraction. Multiple prompting strategies were explored, including prefix, cloze, anticipatory, chain-of-thought, and heuristic prompting.

Table 7: Inter-annotator agreement stratified by cancer type.

Cancer Type	Field	κ	%	Interpretation
Breast	GRADE	0.975	98.83	Almost perfect
	STAGE	0.944	95.48	Almost perfect
	MORPHOLOGY	0.914	96.47	Almost perfect
	T	0.916	94.96	Almost perfect
	N	0.933	95.88	Almost perfect
	M	0.930	95.88	Almost perfect
	LATERALITY	0.958	97.50	Almost perfect
Colorectal	GRADE	0.852	91.49	Almost perfect
	STAGE	0.889	91.49	Almost perfect
	MORPHOLOGY	0.964	97.87	Almost perfect
	T	0.576	73.49	Moderate
	N	0.918	92.42	Almost perfect
	M	0.784	85.38	Substantial
	LATERALITY	0.854	91.49	Almost perfect
Prostate	GRADE	0.992	99.03	Almost perfect
	STAGE	0.983	98.06	Almost perfect
	MORPHOLOGY	0.986	98.71	Almost perfect
	T	0.879	87.74	Almost perfect
	N	0.903	90.32	Almost perfect
	M	0.735	80.65	Substantial
	LATERALITY	0.873	90.32	Almost perfect
Thyroid	GRADE	0.934	97.80	Almost perfect
	STAGE	0.416	24.53	Moderate
	MORPHOLOGY	0.979	99.72	Almost perfect
	T	0.702	71.49	Substantial
	N	0.255	34.84	Fair
	M	0.512	70.57	Moderate
	LATERALITY	0.701	86.88	Substantial

C.1 Grade Extraction Prompts

Table 8: Prompt designs for grade extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract only the numerical tumor grade from the medical report and return it in the JSON template. Ignore additional details. If multiple grades are mentioned, select the final diagnostic grade from the *Impression* section when available. Use the highest numerical tumor grade from the most recent relevant pathology-based report (e.g., pathology or biopsy) and ignore non-pathology reports.
Cloze (p2)	The tumor grade in the text is ___, the grading system used (if available) is ___, and the individual score components (if available) are ___. From the following concatenated clinical notes, extract the highest tumor grade mentioned across all relevant reports. Ignore unrelated or irrelevant sections. The highest tumor grade in the text is ___, the grading system used (if available) is ___, and the individual score components (if available) are ___.
Anticipatory (p3)	[Input 1:] How can tumor grade information be identified in a medical report? [Input 2:] For each tumor mentioned in the given medical report, extract the reported grade. If multiple grades are mentioned, select the final diagnostic grade from the *Impression* section and return only the numerical grade. [Input 1:] How can tumor grade information be identified in a medical history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the highest relevant tumor grade from the most recent pathology report. If multiple grades are mentioned in that report, select the final diagnostic grade from the *Impression* or equivalent summary section. Ignore irrelevant or outdated reports and return only the numerical grade.
Chain-of-Thought (p4)	**EXAMPLE:** Identify and extract the tumor grade from the medical report. If the grading system and its components are mentioned, include them. **TEXT:** "Microscopic examination reveals an invasive ductal carcinoma with moderate tubule formation (score = 2), marked pleomorphism (score = 3), and 5 mitoses per 10 high-power fields (score = 2). Based on these features, the tumor is classified as Nottingham Grade II." **QUESTION 1:** What is the tumor grade mentioned in the medical report? **ANSWER 1:** The tumor grade is II (explicitly stated as Nottingham Grade II). **QUESTION 2:** What grading system is used in the report, if available? **ANSWER 2:** The grading system used is Nottingham. **QUESTION 3:** Is there any additional information about the tumor that should be noted? **ANSWER 3:** No additional features such as necrosis were mentioned. **QUESTION:** Using the stored example, extract the tumor grade, grading system, and score components from the following medical report. **TEXT:** `<text>`
Heuristic (p5)	[Input 1:] First, store these rules in memory: If the tumor grade is explicitly mentioned with a number, use that as the grade. If the grading system is Nottingham, check for component scores such as tubule formation, pleomorphism, and mitotic count. If the grade is described as "Grade I," "Grade II," or "Grade III," classify it accordingly. If it is described as "low," "moderate," or "high," map these to Grade I, II, or III, respectively. If the grading system is not mentioned but descriptors such as "well differentiated," "moderately differentiated," or "poorly differentiated" are used, assign Grade I, II, or III based on those terms. If the grade remains ambiguous, infer it from contextual clues. [Input 2:] Given the patient's full medical record history, extract the tumor grade, grading system, and individual component scores, if available, from the most recent pathology report that discusses tumor grading. Ignore irrelevant or outdated reports without pathology-based grading. Extract: (i) the highest relevant tumor grade in the most recent pathology document, (ii) the grading system used, and (iii) component scores such as tubule formation, pleomorphism, and mitotic count, when available.

C.2 Morphology Extraction Prompts

Table 9: Prompt designs for morphology extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the most relevant morphology from the latest pathology report in the patient's medical history (e.g., pathology or biopsy). If available, prioritize the final morphology stated in the *Impression* or *Diagnosis* section. Select the final answer from the predefined label list: Adenocarcinoma, NOS; Carcinoma, NOS; Follicular adenocarcinoma, NOS; Infiltrating duct carcinoma, NOS; Lobular carcinoma, NOS; Medullary carcinoma, NOS; Mucinous adenocarcinoma; Neoplasm, malignant; Signet ring cell carcinoma. Ignore non-pathology reports. Return only the result in the specified JSON format without additional details.
Cloze (p2)	From the following concatenated clinical notes, extract the most relevant morphology from the latest pathology report. Ignore unrelated or irrelevant sections. The morphology mentioned in the text is ___, and the specific carcinoma type (if available) is ___. Select the final morphology from the predefined label list and return only the result in the specified JSON format.
Anticipatory (p3)	[Input 1:] How can morphology information be identified in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the most relevant morphology reported in the most recent pathology report. If multiple morphology types are mentioned, select the most relevant one. Prioritize the final morphology stated in the *Impression* or equivalent summary section. Ignore irrelevant or outdated reports. Choose the final answer from the predefined label list and return only the result in JSON format.
Chain-of-Thought (p4)	**EXAMPLE:** Identify and extract the morphology from the most recent pathology report. If multiple morphology types are mentioned, select the most relevant one, prioritizing the final diagnosis. **TEXT:** "The pathology report reveals features consistent with invasive ductal carcinoma. The diagnosis concludes invasive ductal carcinoma, NOS." **QUESTION 1:** What morphology is mentioned? **ANSWER 1:** Infiltrating duct carcinoma, NOS. **QUESTION 2:** Are there additional morphology types? **ANSWER 2:** No additional morphology types are introduced. **QUESTION 3:** What is the final morphology? **ANSWER 3:** Infiltrating duct carcinoma, NOS. **TEXT:** "Microscopic examination reveals mucinous adenocarcinoma with signet ring features. The final diagnosis is mucinous adenocarcinoma." **QUESTION 1:** What morphology is mentioned? **ANSWER 1:** Mucinous adenocarcinoma. **QUESTION 2:** Are additional patterns mentioned? **ANSWER 2:** Yes, signet ring features are mentioned, but not as the final diagnosis. **QUESTION 3:** What is the final morphology? **ANSWER 3:** Mucinous adenocarcinoma. Choose the final morphology from the predefined label list.
Heuristic (p5)	[Input 1:] First, apply the following rules: If morphology is explicitly stated, use it directly. If descriptors such as "ductal," "lobular," or "invasive" appear, map them to the corresponding morphology category. Terms such as "mucinous" map to Mucinous adenocarcinoma, and "signet ring" maps to Signet ring cell carcinoma. If only "adenocarcinoma" is mentioned without specification, classify as Adenocarcinoma, NOS. If only "carcinoma" is mentioned, classify as Carcinoma, NOS. Ambiguous descriptions should be mapped to Neoplasm, malignant. [Input 2:] From the patient's full medical history, extract the morphology from the most recent relevant pathology report. Ignore irrelevant or outdated reports. Select one label from the predefined list and return only the result in JSON format.

C.3 TNM Extraction Prompts

Table 10: Prompt designs for TNM staging extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the most relevant TNM staging information from the most recent pathology or oncology report that mentions tumor staging. For each component (T, N, and M), extract the appropriate stage based on the findings. If substages (like T2bN1cM0) are mentioned, only extract the letter and numerical stage associated with each. Ignore non-pathology reports and irrelevant sections. Return the final TNM staging in the JSON template provided. Use the gold standard labels below to fill in each category: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Cloze (p2)	From the following concatenated clinical notes, extract the TNM staging information based on the most relevant pathology report. Ignore unrelated or irrelevant sections. The T stage in the text is ___, the N stage in the text is ___, and the M stage in the text is ___. Ensure that the stages are chosen from the following gold standard labels: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx If substages like T2bN1cM0 are mentioned, extract only the letter and numerical stage for each component. Your final response should be in the given JSON format.
Anticipatory (p3)	[Input 1:] How to identify TNM staging information in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the TNM staging information reported in the most recent pathology report. The TNM staging consists of the primary tumor stage (T), the regional lymph node stage (N), and the distant metastasis stage (M). If any substages are mentioned, extract only the letter and numerical stage associated with each component (e.g., T2bN1cM0 would be T2, N1, M0). Ignore irrelevant or outdated reports. Return the TNM stages in the provided JSON format. T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Chain-of-Thought (p4)	EXAMPLE: Identify and extract the TNM staging from the most recent relevant pathology report. If the staging includes substages (e.g., T2b, N1c), extract only the main stage component (e.g., T2, N1). Use only the values from the gold standard label list. TEXT: "The pathology report states: Tumor shows invasion into muscularis propria consistent with T2b staging. There is evidence of one positive regional lymph node (N1c). No distant metastasis identified (M0)." QUESTION 1: What is the T stage mentioned in the report? ANSWER 1: The report mentions T2b, so the T stage is T2. QUESTION 2: What is the N stage mentioned in the report? ANSWER 2: The report mentions N1c, so the N stage is N1. QUESTION 3: What is the M stage mentioned in the report? ANSWER 3: The report states M0 explicitly, so the M stage is M0. FINAL RESPONSE: T: "T2", N: "N1", M: "M0" Valid Label Options: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx
Heuristic (p5)	[Input 1:] First, store these rules in memory: - If the T, N, or M stage is explicitly mentioned (e.g., "T2b", "N1c", "M0"), extract only the main stage (e.g., T2, N1, M0). - If a range of stages is given (e.g., "T1–T2", "N0–N1"), choose the highest one. - If only descriptive terms are used (e.g., "localized invasion," "no nodal involvement," "distant spread detected"), infer the TNM stage from context: "Localized tumor" → likely T1 or T2 "No lymph node involvement" → N0 "Metastatic disease" → M1 Prioritize the most recent relevant pathology report in the patient's medical history. Ignore outdated or irrelevant notes (e.g., preoperative assessments, non-pathology documentation). Use only the following label options for the final output: T: T0, T1, T2, T3, T4, Tis, Tmic, Tx N: N0, N1, N2, N3, Nx M: M0, M1, Mx [Input 2:] Given the patient's full medical record history, extract the TNM staging by identifying the most recent pathology report that provides clear staging information. Apply the rules above to extract: - T stage (primary tumor) - N stage (regional lymph nodes) - M stage (distant metastasis) Return your final answer in the given JSON format. Only use the values from the predefined label lists. Do not generate your own stages.

C.4 Laterality Extraction Prompts

Table 11: Prompt designs for laterality extraction.

Prompt Type	Prompt Text
Prefix (p1)	Extract the tumor laterality from the latest pathology report in the patient's medical history (e.g., pathology, biopsy). If available, prioritize the final laterality mentioned in the 'Impression' or 'Diagnosis' section. Choose your final answer from the following list: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un Ignore non-pathology reports. Return ONLY in the given JSON format, with a single key-value pair. Do NOT include additional qualifiers, sites, modifiers, or nested fields.
Cloze (p2)	From the following concatenated clinical notes, extract the tumor laterality from the latest pathology report. Ignore unrelated or irrelevant sections. The tumor laterality mentioned in the text is ___. Choose the laterality from the following list, and return ONLY in the given JSON format: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Anticipatory (p3)	[Input 1:] How to identify tumor laterality in a medical report history? [Input 2:] For each tumor mentioned in the patient's medical history, extract the laterality reported in the latest pathology report. If laterality is mentioned in multiple places within the same report, prioritize the clearest and most complete expression, especially from the 'Impression', 'Diagnosis', or equivalent summary sections. Anticipate ambiguities such as paired organs without a side specified, and distinguish between bilateral, unilateral, and unknown cases. Ignore outdated or irrelevant reports. Choose the laterality from the following options, and return ONLY in the given JSON format: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Chain-of-Thought (p4)	INSTRUCTIONS: Identify and extract the laterality from the most recent pathology report. If laterality is mentioned multiple times, prioritize the clearest and most definitive expression, especially from the 'Impression', 'Diagnosis', or equivalent summary sections. TEXT: "The tumor is located in the right breast, with no evidence of contralateral involvement. The final diagnosis section confirms malignancy in the right breast." QUESTION 1: What laterality is mentioned in the medical report? ANSWER 1: The laterality mentioned is right, primary organ (explicitly stated as right breast). QUESTION 2: Are there any ambiguous or conflicting laterality statements elsewhere in the report? ANSWER 2: No conflicting laterality is reported; all references are to the right side. QUESTION 3: What is the final laterality you would assign to this tumor based on the report? ANSWER 3: The final laterality assigned is right, primary organ. TEXT: "Pathology notes indicate bilateral lesions, with suspicious masses observed in both lobes. The diagnosis section confirms bilateral malignancy." QUESTION 1: What laterality is mentioned in the medical report? ANSWER 1: The laterality mentioned is bilateral involvement. QUESTION 2: Are there any ambiguous or conflicting laterality statements elsewhere in the report? ANSWER 2: No, the statement about bilateral malignancy is consistent. QUESTION 3: What is the final laterality you would assign to this tumor based on the report? ANSWER 3: The final laterality assigned is bilateral involvement. Choose the laterality from the following options: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un
Heuristic (p5)	[Input 1:] First, store these rules in memory: - If laterality is explicitly stated as right, left, or bilateral, use that information as the laterality type. - If the tumor is located in a paired organ (e.g., breast, kidney, lung), and the report mentions both sides, classify it as bilateral involvement. - If only one side is referenced (e.g., "mass in the right lobe"), assign left, primary organ or right, primary organ accordingly. - If the site is a paired organ but the laterality is not specified, classify it as paired, lat. unknown. - If the tumor occurs in an organ that is not paired (e.g., stomach, pancreas), classify it as not a paired site/un. [Input 2:] Given the patient's full medical report history, extract the laterality, ensuring you choose one of the gold standard labels. Identify the most recent relevant pathology report that includes a laterality mention. Ignore reports that are outdated or do not include laterality cues relevant to a paired primary tumor site. Use the rules mentioned above to extract: - The laterality of the tumor described in the most recent relevant pathology report. Make sure to choose the laterality from the following list: - bilateral involvement - left, primary organ - right, primary organ - paired, lat. unknown - not a paired site/un

D Multi-Agent Prompt Designs

D.1 Chunking Agent Prompt

Table 12: Prompt design for the chunking agent.

Component	Prompt Text

Prompt	You are a medical NLP system that segments ANY clinical text (even if it has no pathology-style headers) into semantically meaningful chunks.
	Output requirements - Return ONLY a single JSON object (no markdown, no explanations). - Schema (strict):
	`{`
	`  "chunk": [`
	`  {`
	`  "section": "string",`
	`  "text": "string",`
	`  "focus": "grade | morphology | tnm | laterality | treatment | other"`
	`  }`
	`  ]`
	`}`
	- The "chunk" array MUST be non-empty (never return an empty list). - Each item MUST have all three fields. - "focus" must be exactly one of: grade, morphology, tnm, laterality, treatment, other.
	Coverage & missing categories (very important) - Never fabricate text. - If a category (e.g., grade) is completely absent in the input, simply do NOT emit a chunk for that category. Downstream will treat missing categories as None. - The overall output MUST still contain at least one chunk. If no focal info is found, emit a single fallback chunk: - section: "other" - text: a short verbatim excerpt from the input (e.g., the first 1–3 sentences or the most informative line) - focus: "other"
	General rules (works for messy text) - The input may be free text, clinic letters, summaries, MDT notes, radiology-pathology blends, bullet lists, tables, or copy-paste artifacts. - If there are no formal headers, infer a best-fit "section" from this shortlist: ["diagnosis","gross","microscopy","immuno","comment","staging","clinical","treatment","history","other"]. - If the text mixes topics, split into small focused chunks (~1–4 sentences) and assign the primary focus: • grade → tumor grading phrases (e.g., "grade 2", "moderately differentiated", "Gleason group 3") • morphology → histologic type/cell type (e.g., "invasive ductal carcinoma", "adenocarcinoma", "papillary carcinoma") • tnm → any T/N/M mentions, stage group, pathologic vs clinical, even partial like "pT3", "N1", "M0", or "indeterminate" • laterality → left/right/bilateral/unifocal/multifocal; single-site organs can be "not applicable" only if clearly stated • treatment → surgery, chemo, radio, hormonal, targeted, immuno; past or planned • other → demographics, dates, admin, unrelated text - Keep each "text" verbatim (trim only leading/trailing whitespace and bullets). Do NOT invent placeholders like "none" or "[absent]".
	Normalization hints BEFORE deciding focus - TNM appears as "pT3 N1 M0", "T3N1M0", "T3 N1 (Mx)", embedded in sentences, or on separate lines. - Grades: prose ("moderately differentiated"), numerals ("grade II/2"), Gleason groups/scores. - Laterality: "Lt", "Rt", "B/L", "bilat." map to left/right/bilateral. - Morphology often near "diagnosis", "type", "histology", "consistent with".
	Edge cases & guarantees - If you cannot find any obvious clinical focus, still emit the fallback "other" chunk (see Coverage). - If TNM elements are scattered (e.g., T far from N), use multiple tnm chunks. - Prefer short, precise chunks over one giant block.
	Pathology/clinical text: `{report}`

D.2 Retriever Agent Prompt

Table 13: Prompt design for the retriever agent.

Component	Prompt Text

Prompt	You are a medical AI assistant helping extract information from pathology report chunks.
	Goal: Find the chunks most relevant to the field: **{field}**
	Chunks: `{chunks_text}`
	Instructions: - Return only relevant chunks for the field - Return ONLY JSON array of strings - If none → return []
	Example:
	`[`
	`  "Grade 2 tumor with moderate atypia."`
	`]`
	or
	`[]`
	No explanation.

D.3 Grade Extraction Agent Prompt

Table 14: Prompt design for the grade extraction agent.

Component	Prompt Text

Prompt	You are an expert medical assistant specialized in cancer pathology information extraction.
	Task: extract tumor grade.
	Return JSON only:
	1. "grade stated" 2. "grade estimated" (must be from valid values) 3. "grade_CD" 4. "grade_evidence"
	VALID VALUES: `{valid_values_bulleted}`
	Rules: - If no info → Unknown - Confidence 0–1 - Evidence = short quote
	Input: `{context}`
	Output:
	`{`
	`  "grade stated": "...",`
	`  "grade estimated": "...",`
	`  "grade_CD": 0.0,`
	`  "grade_evidence": "..."`
	`}`

D.4 Morphology Extraction Agent Prompt

Table 15: Prompt design for the morphology extraction agent.

Component	Prompt Text
Prompt	You are a medical language model specialized in extracting tumor morphology from pathology reports. Your task is to extract and estimate the tumor **morphology** from the given report sections below. Please output the following four key-value pairs: 1. "morphology stated": <The morphology as written in the pathology report, even if not standardized. If no morphology is directly stated, return "Unknown">, 2. "morphology estimated": <The closest standardized morphology based on the list below. Use your clinical knowledge to resolve synonymy or near matches. If unsure, return "Unknown">, 3. "morphology_CD": <Certainty degree of the estimation between 0.00 and 1.00>, 4. "morphology_evidence": <Key text or rationale supporting your estimation> Valid standardized morphologies: - Adenocarcinoma, NOS - Carcinoma, NOS - Follicular adenocarcinoma, NOS - Infiltrating duct carcinoma, NOS - Lobular carcinoma, NOS - Medullary carcinoma, NOS - Mucinous adenocarcinoma - Neoplasm, malignant - Signet ring cell carcinoma - Unknown Report Excerpt: `{context}` Output Format (JSON Only): `{ "morphology stated": "...", "morphology estimated": "...", "morphology_CD": ..., "morphology_evidence": "..." }`

D.5 TNM Extraction Agent Prompt

Table 16: Prompt design for the TNM extraction agent.

Component	Prompt Text
Prompt	You are a clinical information extraction assistant. Your task is to extract the patient's **TNM stage** from the pathology report excerpt below. Please extract the following key-value pairs for each TNM component: **T (Primary Tumor), N (Lymph Nodes), and M (Metastasis)**. Each component should have the following 4 fields: 1. "<component> stated": <As written in the pathology report — e.g., "T2", "N1", "M0", or "Unknown">, 2. "<component> estimated": <Your best estimate based on AJCC 7th edition, if the stated value is missing or ambiguous. Use "Unknown" if not inferable>, 3. "<component>_CD": <Certainty degree of the estimation from 0.00 to 1.00>, 4. "<component>_evidence": <Supporting text or reasoning from the report> Allowed values: - **T (Primary Tumor)**: ["T0", "Tis", "Tmic", "T1", "T2", "T3", "T4", "Tx", "Unknown"] - **N (Regional Lymph Nodes)**: ["N0", "N1", "N2", "N3", "Nx", "Unknown"] - **M (Distant Metastasis)**: ["M0", "M1", "Mx", "Unknown"] Report Sections: `{context}` Output Format (JSON Only): `{ "T stated": "...", "T estimated": "...", "T_CD": ..., "T_evidence": "...", "N stated": "...", "N estimated": "...", "N_CD": ..., "N_evidence": "...", "M stated": "...", "M estimated": "...", "M_CD": ..., "M_evidence": "..." }`

D.6 Laterality Extraction Agent Prompt

Table 17: Prompt design for the laterality extraction agent.

Component	Prompt Text

Prompt	You are an expert medical assistant specialized in cancer pathology information extraction.
	Your task is to extract and estimate the **tumor laterality** from the provided pathology report excerpt below.
	Please extract the following four key-value pairs related to the **tumor laterality**:
	1. "laterality stated": <The laterality as explicitly mentioned in the pathology report: "left", "right", "bilateral involvement", or "not primary site/unpaired">,
	2. "laterality estimated": <Your best estimate of laterality based on clinical knowledge. Use "not primary site/unpaired" if unclear or unpaired organ>,
	3. "laterality_CD": <Certainty Degree between 0.00 and 1.00 reflecting confidence in the estimation>,
	4. "laterality_evidence": <A short explanation or quote from the input text that supports your estimation>
	Valid laterality values: ["left", "right", "bilateral involvement", "not primary site/unpaired"]
	Input: `{context}`
	Output Format (JSON Only):
	`{`
	`  "laterality stated": "...",`
	`  "laterality estimated": "...",`
	`  "laterality_CD": ...,`
	`  "laterality_evidence": "..."`
	`}`

D.7 Reviewer Agent Prompt

Table 18: Prompt design for the reviewer agent.

Component	Prompt Text

Return specification	`{`
	`  "grade_valid": true or false,`
	`  "morphology_valid": true or false,`
	`  "morphology_mapped": "..." or null,`
	`  "tnm_valid": true or false,`
	`  "laterality_valid": true or false or null,`
	`  "treatment_valid": true or false or null,`
	`  "comment": "<summarize inconsistencies or mapping decisions>"`
	`}`

Prompt	You are a clinical reviewer. Validate and map the extracted fields from a pathology report. Output **ONLY** a single JSON object; no prose, no markdown, no code fences. Input values: - Tumor Grade: `{grade}` - Morphology: `{morphology}` - TNM Stage: `{tnm}` - Laterality: `{laterality}` - Treatment: `{treatment}` Accepted Grade Values: ["Grade I", "Grade II", "Grade III", "Unknown"] Accepted Morphology Values: ["Adenocarcinoma, NOS", "Carcinoma, NOS", "Follicular adenocarcinoma, NOS", "Infiltrating duct carcinoma, NOS", "Lobular carcinoma, NOS", "Medullary carcinoma, NOS", "Mucinous adenocarcinoma", "Neoplasm, malignant", "Signet ring cell carcinoma", "Unknown"] Accepted TNM Components: - T: ["T0", "Tis", "Tmic", "T1", "T2", "T3", "T4", "Tx", "Unknown"] - N: ["N0", "N1", "N2", "N3", "Nx", "Unknown"] - M: ["M0", "M1", "Mx", "Unknown"] Accepted Laterality Values: ["left, primary organ", "not a paired site/un", "paired, lat. unknown", "right, primary organ", "total colon", "unknown"] Accepted Treatment Values: ["chemotherapy", "chemotherapy + radiotherapy", "chemotherapy + surgery", "chemotherapy + surgery + radiotherapy", "chemotherapy+hormonal", "chemotherapy+radiotherapy", "hormonal", "invalid code.", "neoadjuvant chemotherapy+surgery+radiotherapy+hormonal", "radical prostatectomy", "radiotherapy + surgery", "surgery", "surgery+chemotherapy", "surgery+chemotherapy+hormonal", "surgery+chemotherapy+radiotherapy", "surgery+chemotherapy+radiotherapy+hormonal", "surgery+hormonal", "surgery+radiotherapy+hormonal", "total thyroidectomy", "unknown"] Rules: - If you are unsure, use null for $*_v alid fields that can be unknown, or false when it clearly doesn't match. - DO NOT include any explanations; return JSON only.$ Return JSON exactly in this shape: `{return_spec}`

Component	Prompt Text
Repair prompt	You will receive text that SHOULD be a JSON object with the following keys and types: {return_spec} Fix it if needed and return ONLY a valid JSON object. No comments, no markdown, no code fences. Text: {bad}

D.8 Evaluator Agent Prompt

Table 19: Prompt design for the evaluator agent.

Component	Prompt Text

Prompt	You are a clinical NLP evaluator.
	Your task is to estimate the confidence score (from 0 to 100) for the accuracy of each extracted field from a pathology report: - Tumor **grade** - Tumor **morphology** - **TNM stage** - **Laterality** - **Treatment**
	Base your estimate on: - Whether the extracted value is complete and clinically valid - Whether it follows accepted terminology - Whether it is likely justified based on context
	Input Fields: - Grade: {grade} - Morphology: {morphology} - TNM: {tnm} - Laterality: {laterality} - Treatment: {treatment}
	Output Format (JSON Only):
	`{`
	`  "grade_confidence": integer (0-100),`
	`  "morphology_confidence": integer (0-100),`
	`  "tnm_confidence": integer (0-100),`
	`  "laterality_confidence": integer (0-100),`
	`  "treatment_confidence": integer (0-100)`
	`}`

Role in pipeline	The evaluator agent performs two complementary functions: 1. **Ground-truth comparison (deterministic)**: It compares extracted values against reference annotations (if available) and assigns correctness flags: `grade_correct, morphology_correct, tnm_correct, laterality_correct, treatment_correct` 2. **Confidence estimation (LLM-based)**: If ground truth is partially or fully unavailable, the agent invokes the LLM to assign confidence scores (0–100) for each field. This dual strategy ensures: - objective evaluation when labels exist - robust fallback when labels are missing

D.9 Aggregation Agent Prompt

Table 20: Prompt design for the aggregation agent.

Component	Prompt Text
Prompt	You are a clinical summarization assistant. Your task is to compile a structured report using STRICT JSON format. Important: - DO NOT return markdown or explanation. - DO NOT wrap the JSON in "'`json` or any formatting. - Return ONLY valid JSON exactly matching this structure: {json_spec} Input variables: Grade: {grade}, Stated: {grade_stated}, Estimated: {grade_estimated}, Confidence: {grade_confidence}, Valid: {grade_valid}, Correct: {grade_correct}, Evidence: {grade_evidence} Morphology: {morphology}, Stated: {morphology_stated}, Estimated: {morphology_estimated}, Confidence: {morphology_confidence}, Valid: {morphology_valid}, Correct: {morphology_correct}, Evidence: {morphology_evidence} TNM: {tnm}, Confidence: {tnm_confidence}, Valid: {tnm_valid}, Correct: {tnm_correct} T: Stated: {t_stated}, Estimated: {t_estimated}, Evidence: {t_evidence} N: Stated: {n_stated}, Estimated: {n_estimated}, Evidence: {n_evidence} M: Stated: {m_stated}, Estimated: {m_estimated}, Evidence: {m_evidence} Laterality: {laterality}, Stated: {laterality_stated}, Estimated: {laterality_estimated}, Confidence: {laterality_confidence}, Valid: {laterality_valid}, Correct: {laterality_correct}, Evidence: {laterality_evidence} Treatment: {treatment}, Stated: {treatment_stated}, Estimated: {treatment_estimated}, Confidence: {treatment_confidence}, Valid: {treatment_valid}, Correct: {treatment_correct}, Evidence: {treatment_evidence}

Component	Prompt Text
Repair prompt	You will receive a model output that SHOULD be valid JSON matching this schema: `{json_spec}` The text may contain extra words, code fences, or invalid JSON. Return ONLY a corrected JSON object that strictly conforms to the schema above. No markdown, no explanation. Text to fix: `{bad}`

D.10 Matrix Calculator Agent Prompt

Table 21: Prompt design for the matrix calculator agent.

Component	Prompt Text
Role	LLM-only, ground-truth-driven evaluation utility for the Matrix Calculator app. The closed set is built from ground-truth unique labels for the current site × category. The LLM must choose **one** option from a numbered list and return only the corresponding number. For TNM categories, substages are collapsed to parent labels for both ground truth and valid labels. MRN alignment is preserved without duplicate collapsing.
LLM mapping prompt	You are a medical coding normalizer. Category: `{category}` Select the single BEST match for the raw prediction from the options below. OPTIONS (answer with the number only): `{numbered}` RAW PREDICTION: "`{raw_value}`" Rules: - Output ONLY the number (e.g., 2). No words. - Choose the closest semantic match. - Handle synonyms, abbreviations, Roman numerals, punctuation, and casing. - If meaning is unknown or unclear, choose the option labeled `unknown` when available.
Strict retry prompt	You are a medical coding normalizer. Category: `{category}` Select the single BEST match for the raw prediction from the options below. OPTIONS (answer with the number only): `{numbered}` RAW PREDICTION: "`{raw_value}`" Rules: - Output ONLY the number (e.g., 2). No words. - Choose the closest semantic match. - Handle synonyms, abbreviations, Roman numerals, punctuation, and casing. - If meaning is unknown or unclear, choose the option labeled `unknown` when available. IMPORTANT: Respond with ONLY the number of the chosen option. No other text.
Evaluation logic	For each category (`grade`, `morphology`, `t`, `n`, `m`, `laterality`), the agent: - resolves the corresponding prediction and ground-truth columns, - builds the valid closed set from site-specific ground-truth labels, - maps predictions into that closed set using the LLM, - computes classification metrics using `classification_report`, - computes a confusion matrix using `confusion_matrix`.
TNM parent collapse	For TNM categories, substages are collapsed to parent labels before evaluation. Examples include: - `pT1a` → `t1` - `Tis` → `tis` - `N1c` → `n1` - `M0` → `m0` - unknown or empty values → `unknown`

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

Ashwin Kirubakaran
Edison Academy Magnet School
Edison, NJ, USA
ashwinkiru10@gmail.com

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Resolving contradictions in biomedical literature requires more than factual recall; it demands identifying the hidden variables that explain divergent findings. Existing NLI benchmarks such as MedNLI operate at the sentence level and fail to capture document-level conflicts driven by differences in dosage, cell type, or study design. We introduce BioConflict, a benchmark of 250 expert-annotated paper pairs (500 abstracts) across ten biomedical topics, formalizing three tasks: conflict detection, contextual variable extraction, and consensus synthesis. We evaluate five general-purpose large language models and two domain-specific baselines, finding that general-purpose large language models achieve strong conflict detection (F1 up to 0.89) but exhibit brittle reasoning in synthesis, while domain-specific models lag significantly on all generative tasks. These findings highlight the need for context-aware biomedical AI capable of resolving, not merely retrieving, conflicting scientific evidence.

1 Introduction

The validity of scientific conclusions in biomedicine is often contingent on the specific context in which an experiment was conducted. A single molecule can act as both a therapeutic agent and a toxin depending on its concentration, timing of administration, or the genetic background of the host. TGF-beta, for example, functions as a tumor suppressor in early-stage cancer yet drives metastasis in advanced disease. This kind of context-dependence is not an anomaly; it is the norm across molecular biology, pharmacology, and epidemiology. Detecting these contradictions is essential for sound clinical literature review, yet the sheer volume of new publications makes manual synthesis increasingly impractical. PubMed now contains over 36 million entries and receives more than one million new records per year (Lu, 2011; Jin et al., 2024), a rate that makes exhaustive human curation impossible at scale.

Prior BioNLP work has made substantial progress on information extraction. Benchmarks like BLURB (Gu et al., 2022) cover named entity recognition, relation extraction, and question answering, and have driven real advances in domain-specific pretraining. Contradiction detection, however, has largely been treated as a variant of Natural Language Inference (NLI). MedNLI (Romanov and Shivade, 2018) is the most prominent clinical NLI dataset. It is physician-annotated, grounded in patient records from MIMIC-III, and has served as a strong benchmark for biomedical language understanding. MedNLI operates on isolated sentence pairs drawn from clinical notes. Real scientific conflicts, however, are cross-document phenomena. They arise when two full studies, each internally valid, reach incompatible conclusions because they differ in ways that neither abstract makes explicit, such as the cancer stage of the patient cohort, whether the drug formulation was selective or non-selective, or whether dietary exposure was measured in cups per day or grams of caffeine. These distinctions are difficult to capture in sentence-level NLI settings.

The rise of Retrieval-Augmented Generation (Lewis et al., 2020) makes cross-document conflict resolution practically urgent. When a clinical decision support system retrieves two contradictory papers in response to a query, the model must either adjudicate between them or produce a synthesis that explains the discrepancy in terms a clinician can act on. Current evidence suggests that models often absorb retrieved context even when it conflicts with their pretraining knowledge, but they are equally capable of hallucinating confident-sounding justifications for incorrect or unsupported claims (Huang et al., 2025). The "lost-in-the-middle" effect (Liu et al., 2024) compounds this: when relevant information is buried in a long

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 552–558
July 3-4, 2026 ©2026 Association for Computational Linguistics

prompt, models systematically underweight it relative to content near the beginning or end of the context window.

BioConflict is designed to probe all three of these failure modes in a unified benchmark. Our contributions are: (1) a gold-standard corpus of 250 paper pairs spanning ten topics with documented recurring conflicts; (2) three evaluation tasks of increasing cognitive complexity, namely conflict detection, contextual variable extraction, and consensus synthesis; and (3) a systematic evaluation of five general-purpose large language models and two domain-specific models, revealing consistent gaps between classification performance and generative reasoning quality.

2 Methodology

2.1 Topic Selection and Corpus Construction

We selected ten biomedical topics on the basis of two criteria: the topic has generated a substantial peer-reviewed literature with documented conflicting findings, and the conflict can be traced to at least one identifiable hidden variable rather than outright experimental error. The selected topics span three broad areas: molecular signaling pathways (NOTCH, ROS, TGF-beta, Autophagy, p53), pharmacological interventions (Beta-blockers, Metformin, VEGF inhibitors), and epidemiological health claims (Coffee, Vitamin D). For each topic we retrieved 50 papers from PubMed and manually curated 25 opposing pairs, yielding 250 pairs and 500 abstracts in total. Table 1 lists each topic, the nature of its primary conflict, and the key hidden variable that explains the divergence.

Abstracts were retrieved via the NCBI Entrez API using Biopython's Bio.Entrez module (Cock et al., 2009). We used batched fetching of 200 identifiers per call to remain within the 3-requests-per-second rate limit. All content in BioConflict consists of authentic peer-reviewed text; no synthetic or model-generated abstracts were included.

2.2 Benchmark Tasks

BioConflict evaluates models on three tasks of increasing complexity, designed so that failure on a later task cannot be attributed solely to deficiencies measured by an earlier one.

Task 1: Conflict Detection. Given two abstracts, the model classifies their relationship into one of six categories: Direct Contradiction, Context-dependent (both correct under different conditions), Methodological Difference, Population or Species Difference, Temporal Difference, or Dose-dependent. This taxonomy goes beyond binary NLI entailment. A model must not only recognize that a disagreement exists but correctly identify its scientific basis, which requires integrating information from both abstracts simultaneously rather than evaluating each in isolation.

Task 2: Contextual Variable Extraction. Given the same abstract pair, the model must extract the specific hidden variable that explains why the two papers reached different conclusions. Acceptable extractions include named entities such as cancer stage, drug selectivity, dosage range, cell line, or study design type. Scoring uses a strict binary criterion in which an extraction is correct only if the value, unit, and any relevant time-point all match the expert-verified label exactly. Partial credit is not given.

Task 3: Consensus Synthesis. The most demanding task requires the model to generate a "Golden Consensus" statement that reconciles the apparent conflict by explicitly invoking the hidden variable from Task 2. A correct synthesis for the TGF-beta topic, for instance, must articulate that TGF-beta suppresses tumor growth in early-stage carcinomas via the canonical SMAD pathway, but promotes epithelial-mesenchymal transition and metastasis in advanced tumors through SMAD-independent signaling. Outputs are evaluated using the G-Eval framework (Liu et al., 2023), in which GPT-o3-mini acts as an LLM judge with chain-of-thought prompting and scores each synthesis on factuality, resolution quality, and clinical grounding on a 1–5 Likert scale.

2.3 Annotation Protocol and Model Selection

Ground truth for all three tasks was established through a multi-annotator expert annotation process using a custom interface built in Google Colab. Multiple annotators independently annotated each paper pair, recording a free-text description of the conflict, identifying the hidden variable with its exact value and units where applicable, and writing a Golden Consensus statement. Disagreements were resolved through adjudication, where annotators reviewed conflicting labels and reached a final consensus after joint discussion. Cases lacking a clearly identifiable conflict, or those containing abstracts of insufficient quality, were excluded prior to inclusion in the final benchmark. Inter-annotator

agreement was assessed on a subset of the data, showing substantial agreement for conflict classification and consistent alignment on variable extraction.

We evaluated two classes of models. Domain-specific baselines were BioBERT (Lee et al., 2020), fine-tuned from BERT-base on PubMed and PMC full texts, and PubMedBERT (Gu et al., 2022), pretrained from scratch on PubMed abstracts with an in-domain vocabulary. This vocabulary avoids fragmenting domain-specific terms such as *acetyltransferase* into semantically meaningless subword pieces, which has been shown to benefit downstream biomedical NLP performance. The general-purpose large language models evaluated were GPT-4o (OpenAI, 2024), Claude 3.5 Sonnet (Anthropic, 2024), Llama 3.1 (70B) (Grattafiori et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and Gemma-3-12B (Gemma Team et al., 2025). All models were evaluated at temperature 0.0 with identical prompts across tasks.

3 Results

Across all three tasks, general-purpose large language models substantially outperformed domain-specific baselines, with the performance gap widening on Tasks 2 and 3. Table 2 reports the complete results.

3.1 Task 1: Conflict Detection

GPT-4o achieved the highest F1 (0.89), followed by Claude 3.5 Sonnet (0.87) and Llama 3.1 (0.85). Gemma-3-12B (0.81) outperformed Qwen2.5-VL-7B (0.79) despite its smaller size. The dominant error was misclassifying methodological differences as direct contradictions, indicating reliance on surface-level disagreement. PubMedBERT (0.74) outperformed BioBERT (0.50), but both struggled with cross-document reasoning, often failing to integrate information across abstracts.

3.2 Task 2: Contextual Variable Extraction

Variable extraction accuracy varied substantially by clinical domain (Table 3). Models performed best in Cardiology and Infectiology, where the key moderating variables are well-standardized across the literature. A model can learn that hypertension status and diabetes comorbidity are typical moderators in cardiovascular studies and perform well without deep mechanistic knowledge. Performance dropped sharply in Oncology and molecular signaling topics, where correctly identifying the hidden variable required pathway-level understanding. In the p53 topic, for example, the distinguishing variable is often the specific mutation type, which determines whether the protein retains partial transcriptional activity; this cannot be inferred from clinical vocabulary patterns alone.

The dominant error type across all large language models (75% of failures) was incorrect unit extraction: a model identified the right variable category but extracted the wrong value or unit, for instance reporting a Vitamin D threshold in ng/mL when the paper used nmol/L. Formula misapplication also appeared frequently in topics requiring threshold comparisons, where models used outdated clinical reference ranges drawn from pretraining data rather than the values stated in the abstract being evaluated.

3.3 Task 3: Consensus Synthesis

GPT-4o and Claude 3.5 Sonnet scored 4.7 and 4.6 respectively, producing syntheses that correctly identified the hidden variable and reconciled both findings. Llama 3.1 scored 4.3, occasionally failing to invoke the hidden variable explicitly even when Task 2 extraction had been correct. Gemma-3-12B (4.0) produced coherent syntheses that identified the conflict type but sometimes lacked clinical specificity. Qwen2.5-VL-7B (3.8) inconsistently grounded reconciliations in the precise values from the abstracts. PubMedBERT and BioBERT scored 2.2 and 2.0 respectively, typically paraphrasing one abstract rather than producing a genuine reconciliation. The GPT-o3-mini judge achieved ICC(2,1) = 0.818 (95% CI: 0.772–0.854) against human raters, with highest agreement on Factuality and lowest on Resolution.

4 Discussion

BioConflict reveals a consistent pattern: models are far more capable at detecting that a conflict exists than at explaining why or synthesizing a resolution. Three failure modes account for most of the gap.

Contextual hallucination. Models sometimes produced mechanistic justifications absent from either abstract. In the beta-blocker topic, one model fabricated a receptor selectivity profile to explain a survival discrepancy (Huang et al., 2025), a confabulation that could directly influence prescribing decisions.

Topic ID	Biological Focus	Nature of Conflict	Key Hidden Variable(s)
NOTCH_dual_role	Notch signaling	Oncogene vs. Tumor Suppressor	Tissue Type / Cell Lineage
ROS_dual_role	Reactive Oxygen Species	Survival vs. Apoptosis	Concentration / ROS Species
TGF-beta_dual	TGF-beta signaling	Suppression vs. Metastasis	Cancer Stage
VEGF_inhibitors	Angiogenesis	Response vs. Resistance	Hypoxia Status / Splicing
autophagy_dual	Cellular recycling	Suppression vs. Promotion	Tumor Microenvironment
beta_blockers	Beta-adrenergic signaling	Improved Survival vs. No Effect	Selective vs. Non-selective
coffee_health	Caffeine consumption	Benefit vs. Risk	Dosage (cups per day)
metformin_cancer	Antidiabetic agent	Prevention vs. No Effect	Study Design (Obs vs. RCT)
p53_context	Tumor protein 53	Apoptosis vs. Senescence	Mutation Type / Stress Type
vitamin_D_cancer	Supplementation	Prevention vs. No Effect	Baseline Deficiency Level

Table 1: BioConflict topics, conflicts, and key hidden variables.

Model	Task 1: Conflict Detection (F1)	Task 2: Variable Extraction (Acc)	Task 3: Consensus Synthesis (G-Eval)
GPT-4o	0.89	0.84	4.7 / 5.0
Claude 3.5 Sonnet	0.87	0.81	4.6 / 5.0
Llama 3.1 (70B)	0.85	0.78	4.3 / 5.0
Qwen2.5-VL-7B	0.79	0.74	3.8 / 5.0
Gemma-3-12B	0.81	0.76	4.0 / 5.0
PubMedBERT	0.74	0.62	2.2 / 5.0
BioBERT	0.50	0.54	2.0 / 5.0
Human Expert	0.98	0.95	4.9 / 5.0

Table 2: Performance across all three BioConflict tasks. G-Eval scores range from 1–5.

Lost-in-the-middle. Models underweighted information in the middle of long prompts (Liu et al., 2024), which is particularly harmful here since the determining variable typically sits in the Methods section, the interior of any combined prompt.

Numerical imprecision. Conflicts often hinge on whether a value crosses a clinical threshold. LLMs misapplied unit conversions or used outdated reference ranges from pretraining; agentic pipelines with code execution can largely eliminate this (Lewis et al., 2020).

Domain pretraining and scale. PubMedBERT's in-domain vocabulary helps on NER (Gu et al., 2022) but does not extend to cross-document reasoning. General-purpose large language models outperformed both domain-specific baselines by large margins, and open-source 70B models (Llama 3.1, Qwen2.5) close most of the gap on Tasks 1 and 2; Task 3 synthesis is where the largest gap remains.

5 Conclusion

BioConflict establishes a new evaluation standard for biomedical conflict reasoning, moving beyond sentence-level NLI to contradiction resolution and consensus synthesis. Our results show that general-purpose large language models detect conflicts reliably but fall short of generating clinically sound reconciliations.

Limitations

The use of abstracts omits methodological detail that sometimes fully explains a conflict. The 250-pair sample may underrepresent ambiguous or emerging disputes, as topics were chosen partly because their conflicts are well-characterized. The LLM-as-judge framework is susceptible to verbosity and self-preference biases (Liu et al., 2023).

References

Anthropic. Claude 3.5 Sonnet model card addendum. Technical report, Anthropic, 2024. URL https://www.anthropic.com/model-card.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. doi: 10.1093/bioinformatics/btp163.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al.

Clinical Domain	GPT-4o (Acc)	Llama 3.1 (Acc)	PubMedBERT (Acc)
Cardiology	0.86	0.82	0.75
Infectiology	0.85	0.81	0.72
Endocrinology	0.84	0.79	0.69
Oncology	0.82	0.74	0.58

Table 3: Variable extraction accuracy by clinical domain for selected models. Full per-model results are consistent with the trends in Table 2.

Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025. doi: 10.1145/3703155.

Q. Jin, R. Leaman, and Z. Lu. PubMed and beyond: biomedical literature search in the age of artificial intelligence. *eBioMedicine*, 100:104988, 2024. doi: 10.1016/j.ebiom.2024.104988.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.153/.

Zhiyong Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011. doi: 10.1093/database/baq036.

OpenAI. GPT-4o system card. Technical report, OpenAI, 2024. URL https://arxiv.org/abs/2410.21276.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187.

A Reliability Metrics

Inter-rater reliability between the LLM judge and human expert annotations is quantified using Cohen's kappa (κ) and the Intraclass Correlation Coefficient ICC(2,1), which treats raters as random effects in a two-way model:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o is the observed proportion of agreement and p_e is the expected proportion under chance. Values of $\kappa > 0.8$ are conventionally interpreted as near-perfect agreement. For continuous Likert scores, ICC(2,1) is preferred; values above 0.75 indicate good reliability. GPT-o3-mini achieved ICC = 0.818 (95% CI: 0.772–0.854). A model judge is flagged for lenience bias if $P(\text{model score} > \text{human score}) \geq 0.95$ across the evaluation set; no flagging occurred for GPT-o3-mini in our Task 3 evaluation.

B Task 1 Prompt: Conflict Detection

The following prompt was used for all Task 1 conflict detection evaluations. All models were queried at temperature 0.0 with this identical prompt. No few-shot examples were provided; all models were evaluated in a zero-shot setting.

```
You are an expert biomedical
researcher tasked with
classifying the relationship
between two scientific abstracts.
You will be given two abstracts
(Paper A and Paper B) from
the peer-reviewed biomedical
literature.  Your task is to
classify their relationship into
exactly one of the following six
categories:

1.  Direct Contradiction:  The
two papers report mutually
incompatible findings with
no contextual qualifier that
reconciles them.

2.  Context-dependent:  Both
papers are correct, but under
different conditions (e.g.,
different tissue types, cancer
stages, or experimental models).

3.  Methodological Difference:
The divergence in findings is
attributable to differences in
study design, assay type, or
measurement approach.

4.  Population or Species
Difference:  The findings
differ because the studies
were conducted in different
populations, species, or cell
lines.

5.  Temporal Difference:  The
findings differ because the
studies examined different time
points or disease stages.

6.  Dose-dependent:  The
findings differ because the
studies used different doses,
concentrations, or exposure
levels of the agent under
investigation.

Read both abstracts carefully
before classifying.  Your
classification must reflect
the underlying scientific
basis for the divergence,
not surface-level lexical
disagreement.

Paper A: {abstract_a}

Paper B: {abstract_b}

Return only valid JSON:
{"label":  "<one of the
six category names above>",
"rationale":  "one or two
sentences"}
```

C Task 2 Prompt: Contextual Variable Extraction

The following prompt was used for all Task 2 contextual variable extraction evaluations. All models were queried at temperature 0.0 with this identical prompt in a zero-shot setting. Scoring was applied post-hoc using a strict binary criterion: an extraction was marked correct only if the variable category, value, unit, and any relevant time-point all matched the expert-verified label exactly. Partial credit was not given.

```
You are an expert biomedical
researcher.  You will be given
two abstracts (Paper A and
Paper B) that reach different
conclusions on the same topic.

Your task is to identify the
single hidden variable that
best explains why the two papers
reached different conclusions.
The hidden variable must be
extracted directly from the text
of the abstracts.  Do not infer
or introduce information that is
not explicitly stated in either
abstract.

Report the hidden variable
with all of the following
fields:  (1) the variable
category (e.g., cancer stage,
drug selectivity, dosage, cell
line, study design); (2) the
exact value or descriptor as
stated in each abstract; (3)
the unit of measurement where
applicable; and (4) the relevant
time-point if the variable is
time-dependent.

Paper A: {abstract_a}

Paper B: {abstract_b}

Return only valid JSON:
{"variable_category":  "string",
"value_paper_a":  "string",
"value_paper_b":  "string",
"unit":  "string or null",
"time_point":  "string or null"}
```

D Task 3 Prompt: LLM Judge (G-Eval)

The following prompt was used for all Task 3 G-Eval evaluations. The hidden variable label from Task 2 annotation was inserted programmatically at the `[VARIABLE]` placeholder.

```
You are an expert medical judge
evaluating a consensus statement
generated from two contradictory
biomedical abstracts.

Criteria:  (1) Factuality:
the consensus contains no
claims unsupported by either
abstract.  (2) Resolution:  the
consensus correctly explains
the disagreement by invoking
the hidden variable.  (3)
Clinical Precision:  the
language is specific enough
to be actionable.

Steps:  State the primary
conclusion of Paper A. State
```

```
the primary conclusion of Paper
B. Identify the hidden variable
[VARIABLE]. Assess whether the
generated consensus correctly
uses this variable to bridge the
two conclusions.

Assign a score from 1 to 5.
Score 5 indicates a synthesis
indistinguishable from a
board-certified clinician's
review.

Return only valid JSON:
{"score":  int, "rationale":
"one or two sentences"}
```

Tokenization Granularity and Medical Term Representations in Language Models

Vojtěch Lanz and Pavel Pecina
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{lanz,pecina}@ufal.mff.cuni.cz

Abstract

We investigate how tokenization granularity affects the representation of medical terminology in language models. Prior work links tokenization granularity to downstream performance under contextualized settings for specifically pretrained and fine-tuned models. We instead ask whether this relationship already emerges at the level of isolated term representations across existing pretrained models. We introduce an intrinsic definition retrieval task using UMLS term-definition pairs, with comparison to WordNet. We show that despite substantially heavier fragmentation of medical terminology, the models remain relatively robust in maintaining semantic alignment between medical terms and their definitions. At the same time, tokenization granularity still correlates with retrieval performance, indicating that effects previously observed in downstream biomedical tasks are already reflected at the level of isolated term representations. Encoder models benefit primarily from whole-token preservation, while for decoder LLMs, tokenization effects emerge mainly at deeper retrieval ranks.

1 Introduction

Tokenization is the first processing step for language models and directly shapes their input representations, thereby influencing overall performance. In practice, tokenizers are typically trained on general-domain corpora (Grattafiori et al., 2024; Jiang et al., 2023; Team et al., 2025; Abdin et al., 2024) and reused when adapting models to specialized domains (Christophe et al., 2024; Labrak et al., 2024b; Sellergren et al., 2025; Corbeil et al., 2025). This mismatch is particularly pronounced in the clinical and biomedical setting, where terminology often differs substantially from general language. As a result, medical terms are frequently segmented into subword units that carry little or no standalone semantic or morphological meaning (Cruz Díaz and Maña López, 2015; Jimenez Gutierrez et al., 2023).

Previous work suggests that contemporary language models are relatively robust to suboptimal tokenization, eventually relying on contextual cues to recover the meaning of poorly segmented clinical terms (Jimenez Gutierrez et al., 2023; Jeong et al., 2023). However, other studies report modest improvements from more suitable segmentation (Ashfaq et al., 2025; Xie, 2026; Jeong et al., 2023). At the same time, recent findings indicate that tokenization granularity can correlate with downstream task performance, particularly in encoder-based architectures trained for domain-specific biomedical applications (Labrak et al., 2024a).

In this work, we revisit this question from a different perspective. Rather than evaluating performance on standard downstream tasks that provide rich contextual signals, we focus on the ability of models to represent medical terms in isolation. We investigate whether tokenization granularity correlates with how effectively models encode medical terminology in their embedding spaces. The emphasis is on the consistency and stability of these representations in semantic alignment and embedding-space proximity, and whether models preserve a coherent representation of a medical concept despite differences in tokenization granularity. We conduct a comparison across publicly available models, including both encoder and decoder architectures, to assess their behavior under inefficient tokenization of medical terms.

To this end, we use UMLS term–definition pairs (Bodenreider, 2004) and frame the task as definition retrieval, where a model must associate a medical term with its correct definition. This setup allows us to assess whether the embedding of a term is relatively closer to its correct definition than to semantically related but distinct ones, avoiding reliance on absolute similarity scores, which could otherwise place many medical concepts close to-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 559–571
July 3-4, 2026 ©2026 Association for Computational Linguistics

	UMLS	WordNet
term–definition pairs	289,531	87,379
words per term	4.19	1.37
words per definition	31.05	10.41

Table 1: Dataset statistics for UMLS and WordNet, reporting number of final term–definition pairs and average lengths of terms and definitions in words.

gether and obscure meaningful differences. We then analyze how performance on this task correlates with tokenizer granularity across models. To provide a general-domain reference point for interpretation, we include a parallel evaluation based on WordNet term–definition pairs (Miller, 1995). The complete source code is available at GitHub[1].

2 Setup

2.1 Data

For our experiments, we use term–definition pairs extracted from UMLS. For comparison, we also consider a general-domain term–definition dataset derived from WordNet. To ensure comparability across models, including those primarily designed for English, we restrict the data to English terms with English definitions.

To maintain a fair evaluation setup for the definition retrieval task, we retain only one term per semantic concept. Specifically, in UMLS we select a single term per concept unique identifier (CUI), and in WordNet we select a single term per synset. To avoid ambiguity arising from multiple surface forms with different senses, we remove duplicate term occurrences.

The final datasets consist of hundreds of thousands of pairs for UMLS and tens of thousands of pairs for WordNet. Table 1 reports the number of final pairs as well as the average number of words per term and per definition. We observe that both terms and definitions in UMLS are approximately three times longer than those in WordNet.

2.2 Tokenization Granularity

To measure how well tokenization preserves word integrity, or conversely how much it fragments words into subword units, we use two metrics: average number of tokens per word (*Avg tokens per word*) and the percentage of out-of-vocabulary words (*OOV words (%)*). OOV words (%) is computed as the proportion of words that are split into more than one token by the tokenizer.

2.3 Definition Retrieval Task

The retrieval task aims to find the most relevant definition for a given term based on cosine similarity among all available definitions. To evaluate performance, we report the *Success@k* curve ($k \in 1, 10, 100, 1000, 2000$), which measures the percentage of queries where the correct definition is within the top k retrieved results.

2.4 Models

We evaluate encoder-based models and decoder large language models, covering general-purpose, biomedical, and multilingual pretraining regimes.

Encoders include general-domain models such as BERT (cased and uncased) (Devlin et al., 2019) and ModernBERT (Warner et al., 2025), medical models such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and Clinical ModernBERT (Lee et al., 2025), as well as multilingual models including mBERT, Distil mBERT (Devlin et al., 2019), mmBERT (Marone et al., 2025), and XLM-R (base and large) (Conneau et al., 2019).

Decoders include general instruction-tuned LLMs such as LLaMA3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Phi-3.5 Mini (Abdin et al., 2024), and Gemma3 (Team et al., 2025), as well as biomedical variants such as BioMistral (Labrak et al., 2024b), Med42 (Christophe et al., 2024), MediPhi (Corbeil et al., 2025), and MedGemma (Sellergren et al., 2025).

For all models, we extract representations from the final hidden layer and apply mean pooling over token embeddings to obtain a single vector. We use this representation for both encoders and decoders to ensure consistency.

3 Results and Analysis

3.1 Tokenization Granularity Across Domains

Building on our evaluation setup, we first analyze how tokenization granularity varies across domains and model families.

We analyze UMLS terms and their definitions separately and compare them with their general-domain counterpart based on WordNet in Table 2, reporting averages across all models (Section 2.4).

The UMLS terms are consistently segmented into more fine-grained subword units than their cor-

[1] https://github.com/lanzv/medical-tokenization-granularity

	UMLS		WordNet		Full Text Tokenization				
	Terms	*Defs*	*Terms*	*Defs*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
Avg tokens per word	2.53	1.79	2.31	1.32	1.41	1.57	1.36	1.80	2.38
OOV words (%)	53.04	31.33	61.13	18.09	24.20	26.83	17.37	33.83	48.74

Table 2: Tokenization statistics averaged across all encoder and decoder models for UMLS and WordNet (terms vs. definitions) and general corpora, including BookSum chapters, Wikipedia pages, PubMed articles, emrQA clinical reports, and Stroke Discharge Summaries (SDS) from UK hospitals.

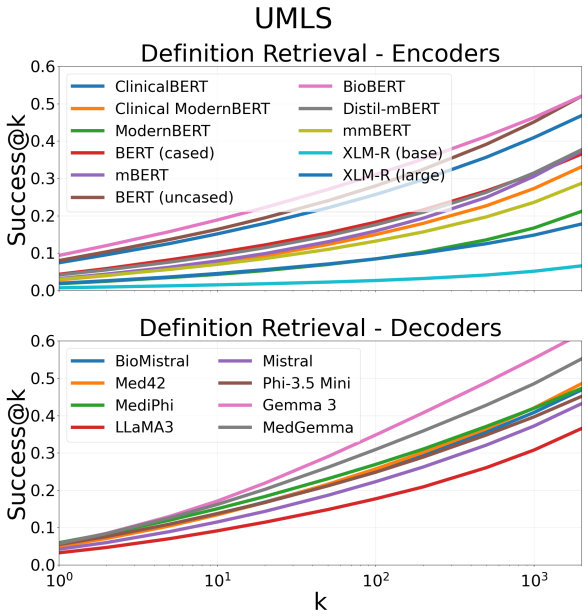


Figure 1: Definition retrieval performance on UMLS measured by Success@k. Encoder models are shown on the top and decoder models on the bottom.

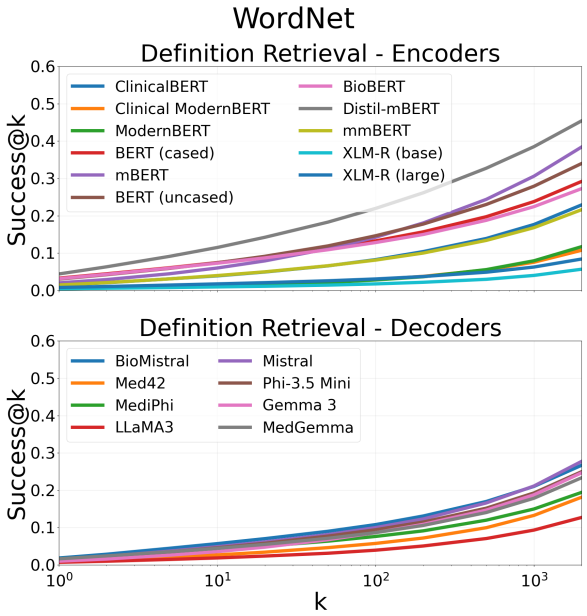


Figure 2: Definition retrieval performance on WordNet measured by Success@k. Encoder models are shown on the top and decoder models on the bottom.

responding definitions. A similar effect is observed in WordNet, indicating that lexical specificity also plays a role in general-domain resources. This consistent gap between terms and definitions suggests that isolated terminological expressions are generally more difficult to represent in subword-based tokenization schemes than more descriptive, sentence-level text, even when the latter is drawn from the same domain and contains many common lexical items. However, fragmentation is most pronounced in UMLS, suggesting higher terminological density in the medical domain.

OOV rates show a different pattern: WordNet terms have slightly higher OOV rates than UMLS. As shown in Table 1, UMLS terms are three times longer, which increases the likelihood of including vocabulary-friendly components such as numbers or stopwords, likely explaining this difference.

To assess how representative the term–definition results are in relation to larger and more practical text sources, we extend the analysis to document-level corpora. We evaluate tokenization on emrQA clinical reports (Pampari et al., 2018), BookSum chapters (Kryscinski et al., 2022), Wikipedia (Foundation), and PubMed articles (Cohan et al., 2018). To control for corpus size, we sample PubMed, BookSum, and Wikipedia documents to match the number of words in emrQA. Tokenization in BookSum and PubMed closely matches WordNet definitions, while Wikipedia shows slightly higher fragmentation. In contrast, emrQA clinical reports align more closely with UMLS definitions. We further analyze a specialized set of 600 stroke discharge summaries from UK hospitals (SDS), which shows fragmentation approaching UMLS-level values, highlighting the effect of clinical specificity on subword segmentation.

A detailed per-model breakdown is provided in Appendix A, along with qualitative tokenization examples in Appendix B.

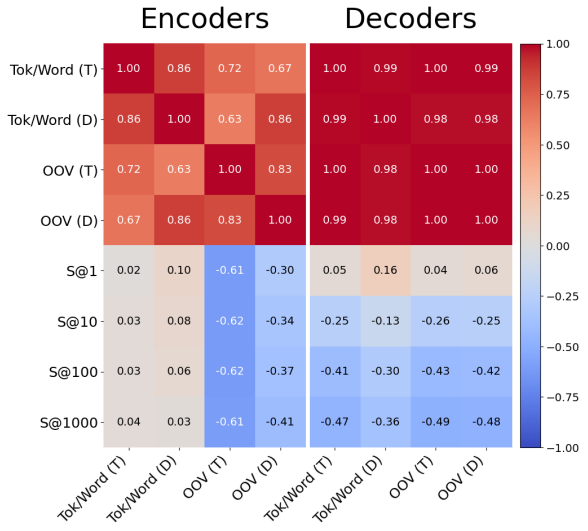


Figure 3: Pearson correlation matrices for UMLS models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (Success@k, denoted as S@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

3.2 Definition Retrieval Performance

We evaluate how well term representations align with their corresponding definitions using a definition retrieval task, which measures the robustness of the embedding space to semantic similarity of the same medical concept represented in different forms. Figures 1 and 2 report Success@k curves for encoder and decoder models on UMLS and WordNet term–definition pairs, allowing a direct comparison of models in terms of retrieval performance and the strength of their semantic alignment.

On UMLS, clinically pre-trained encoders consistently outperform general-domain and multilingual models. This advantage is less pronounced for decoder models, which nevertheless show relatively stable retrieval behavior across architectures.

In contrast, performance on WordNet decreases, despite the smaller candidate pool making the task easier. Drop is more evident for decoder models, which exhibit weaker discrimination in this setting.

These results suggest that models are well adapted to the medical domain, even though tokenization is more fragmented for UMLS compared to WordNet. This indicates a degree of robustness of models to tokenization granularity in semantic representation and alignment of concepts.

3.3 Correlation Between Tokenization and Definition Retrieval

To investigate whether tokenization granularity is related to how well models embed medical terms, we analyze UMLS pairs and measure the correlation between tokenization metrics and the ability of models to place term representations close to their correct definitions in the embedding space. We report Pearson correlations in Figure 3 for encoder and decoder models separately. This allows us to assess whether a consistent trend exists across models, i.e., whether models that split words into more subword units tend to exhibit weaker alignment between terms and their corresponding definitions.

From the results, we observe that for encoder models, the average number of tokens per word (Tok/Word) shows little to no correlation with retrieval performance. In contrast, the proportion of words represented as a single token (OOV) in both definitions and, in particular, terms exhibits a strong correlation with performance across all values of k in Success@k. For decoder models, performance begins to correlate more strongly with tokenization metrics only at higher values of k (i.e., $k = 100$ and $k = 1000$), where both Tok/Word and OOV rate become relevant.

An additional observation is that, for decoder models, tokenization metrics are almost perfectly correlated between terms and definitions, whereas encoders show greater variation.

Overall, the results indicate that tokenization granularity is reflected in how models represent medical terms in their vector space. For encoder models, retrieval performance is largely insensitive to how finely words are split once tokenized; the key factor is whether a word is preserved as a single token. For decoder models, tokenization has little effect on the proximity of term and definition representations, but its influence becomes more apparent when considering a broader neighborhood of nearby representations.

4 Discussion

Although we use definition retrieval as an intrinsic evaluation, it provides a direct and transparent way to assess how a model encodes medical concepts in its embedding space without additional training. It allows us to measure whether term embeddings are consistently and meaningfully aligned with term definitions and how effectively and consistently the model internalizes and represents these concepts.

Although tokenization granularity correlates with retrieval performance, this does not directly indicate its impact on downstream tasks. While previous work has studied such effects in downstream biomedical settings (Labrak et al., 2024a), the general significance and generality of this relationship remain unclear and might vary between tasks.

In addition, we find that English-domain-specific encoders outperform general-domain multilingual models for medical terminology, despite its strong Latin and Greek origins (Jimenez Gutierrez et al., 2023). This contrasts with the findings in span-based clinical question answering (Lanz and Pecina, 2025), where multilingual models tend to perform better. This discrepancy suggests that improvements in multilingual models in downstream question-answering tasks are likely driven by their stronger ability to model contextual interactions and relational structure, rather than by the superior quality of isolated term representations in the embedding space of encoder models. In contrast, for decoder models, no clear advantage of domain-specific pretraining is observed in terms of consistent semantic representation quality.

5 Conclusion

We studied how tokenization granularity relates to the ability of language models to represent medical terms without contextual support, using a definition retrieval task built on UMLS and a WordNet comparison. Our results show that medical terms are more fragmented than general-domain ones, yet models remain largely robust. However, tokenization still plays a role: for encoders, performance depends mainly on whether terms are preserved as single tokens, while for decoders, its effect appears at larger retrieval depths. Overall, tokenization does not fully determine performance, but it influences how medical terms are structured in the embedding space.

Limitations

Our analysis focuses on intrinsic representation properties and does not evaluate downstream task performance; thus, it does not directly capture effects previously observed in task-based studies of tokenization granularity. While definition retrieval provides a controlled proxy for assessing term–definition alignment in embedding space, it remains an indirect measure of representation quality. Finally, our experiments are limited to UMLS-derived term–definition pairs, which restricts coverage to curated medical terminology and may not reflect broader medical and clinical language.

Acknowledgments

This research was partially supported by the SVV project number 260 821 and the Charles University GAUK grant No. 284125. It has also received support and funding from the European Union's Horizon Europe research and innovation programme project *RES-Q plus* (Grant Agreement No. 101057603). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Farzeen Ashfaq, NZ Jhanjhi, Navid Ali Khan, Danish Javed, Mehedi Masud, and Mohammad Shorfuzzaman. 2025. Enhancing ecg report generation with domain-specific tokenization for improved medical nlp accuracy. *IEEE Access*, 13:85493–85506.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *Preprint*, arXiv:2408.06142.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.

Noa P. Cruz Díaz and Manuel Maña López. 2015. An analysis of biomedical tokenization: Problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Geunyeong Jeong, Juoh Sun, Seokwon Jeong, Hyunjin Shin, and Harksoo Kim. 2023. Improving automatic KCD coding: Introducing the KoDAK and an optimized tokenization method for Korean clinical documents. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 96–101, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang and 1 others. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bernal Jimenez Gutierrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362, Toronto, Canada. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickael Rouvier, and Richard Dufour. 2024a. How important is tokenization in French medical masked language models? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8223–8234, Torino, Italia. ELRA and ICCL.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024b. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, {Chan Ho} So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Publisher Copyright: © 2020 Oxford University Press. All rights reserved.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *Preprint*, arXiv:2504.03964.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Andrew Sellergren and 1 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Gemma Team and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and

Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Wenran Xie. 2026. Optimizing biomedical text processing: A comparative analysis of tokenization methods and context-aware representation learning. In *Advanced Data Mining and Applications*, pages 148–160, Singapore. Springer Nature Singapore.

A Tokenization Granularity

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
ClinicalBERT	2.52	1.82	2.44	1.32	1.36	1.58	1.34	1.63	2.18
Clinical ModernBERT	2.34	1.60	2.39	1.26	1.42	1.52	1.22	1.70	2.34
ModernBERT	2.34	1.60	2.39	1.26	1.42	1.52	1.22	1.70	2.34
BERT (cased)	2.80	1.86	2.36	1.27	1.34	1.47	1.34	1.84	2.39
mBERT	2.66	1.87	2.35	1.37	1.43	1.47	1.37	1.82	2.35
BERT (uncased)	2.33	1.73	2.17	1.23	1.31	1.41	1.27	1.55	2.07
BioBERT	2.52	1.82	2.44	1.32	1.36	1.58	1.34	1.63	2.18
Distil-mBERT	2.66	1.87	2.35	1.37	1.43	1.47	1.37	1.82	2.35
mmBERT	2.06	1.60	1.78	1.17	1.35	1.51	1.30	1.77	2.38
XLM-R (base)	2.58	1.88	2.35	1.46	1.44	1.58	1.46	1.75	2.15
XLM-R (large)	2.58	1.88	2.35	1.46	1.44	1.58	1.46	1.75	2.15

Table 3: Average tokens per word across encoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum chapters, Wikipedia pages, PubMed articles, emrQA clinical notes, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
ClinicalBERT	44.13	29.83	58.65	16.47	22.68	27.52	15.97	30.40	45.11
Clinical ModernBERT	50.33	25.03	66.50	15.12	23.13	24.45	10.55	29.21	47.71
ModernBERT	50.33	25.03	66.50	15.12	23.13	24.45	10.55	29.21	47.71
BERT (cased)	56.25	30.63	55.07	14.11	22.14	23.43	15.97	34.15	50.52
mBERT	58.23	34.87	63.10	21.28	27.02	26.22	19.84	36.73	52.35
BERT (uncased)	40.18	27.96	49.73	12.35	21.26	21.62	13.12	27.18	41.66
BioBERT	44.13	29.83	58.65	16.47	22.68	27.52	15.97	30.40	45.11
Distil-mBERT	58.23	34.87	63.10	21.28	27.02	26.22	19.84	36.73	52.35
mmBERT	37.18	23.64	41.00	9.48	19.92	21.34	10.04	25.54	40.60
XLM-R (base)	63.41	40.81	67.65	29.49	28.72	34.35	33.41	44.43	54.59
XLM-R (large)	63.41	40.81	67.65	29.49	28.72	34.35	33.41	44.43	54.59

Table 4: OOV words (%) across encoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
BioMistral	2.87	1.95	2.42	1.39	1.49	1.75	1.49	2.06	2.77
Med42	2.35	1.65	2.31	1.25	1.36	1.49	1.28	1.67	2.09
MediPhi	2.94	2.01	2.50	1.45	1.53	1.77	1.54	2.13	2.85
LLaMA3	2.35	1.65	2.31	1.25	1.36	1.49	1.28	1.67	2.09
Mistral	2.87	1.95	2.42	1.39	1.49	1.75	1.49	2.06	2.77
Phi-3.5 Mini	2.94	2.01	2.50	1.45	1.53	1.77	1.54	2.13	2.85
Gemma 3	2.14	1.60	2.07	1.19	1.37	1.52	1.29	1.76	2.42
MedGemma	2.14	1.60	2.07	1.19	1.37	1.52	1.29	1.76	2.42

Table 5: Average tokens per word across decoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA clinical notes, and Stroke Discharge Summaries (SDS).

Model	UMLS		WordNet		Full Text Tokenization				
	Term	*Def*	*Term*	*Def*	**BookSum**	**Wikipedia**	**PubMed**	**emrQA**	**SDS**
BioMistral	63.30	36.01	64.01	21.20	25.90	30.64	20.30	38.54	54.91
Med42	48.98	27.69	63.95	14.57	22.29	24.07	12.12	29.80	42.14
MediPhi	67.29	39.03	68.78	25.28	27.52	32.70	23.59	41.94	58.03
LLaMA3	48.98	27.70	63.95	14.56	22.29	24.07	12.12	29.80	42.14
Mistral	63.30	36.01	64.01	21.20	25.90	30.64	20.30	38.54	54.91
Phi-3.5 Mini	67.29	39.03	68.78	25.28	27.52	32.70	23.59	41.94	58.03
Gemma 3	41.41	23.22	55.20	10.49	20.98	21.73	9.71	26.88	41.78
MedGemma	41.41	23.22	55.20	10.49	20.98	21.73	9.71	26.88	41.78

Table 6: OOV words (%) across decoder models for UMLS (terms and definitions), WordNet (terms and definitions), and full-text corpora including BookSum, Wikipedia, PubMed, emrQA, and Stroke Discharge Summaries (SDS).

B Qualitative Tokenization Examples

Model	Infertility	Spectroscopy	Neurochemistry	Excipients	Bacteroides
bert-base-cased	In-fer-tility	S-pect-ros-copy	N-eur-och-em-ist-ry	Ex-ci-pie-nts	Ba-cter-oid-es
bert-base-multilingual-cased	Inf-erti-lity	Sp-ect-ros-co-py	Neu-roch-emi-str-y	Ex-ci-pien-ts	Ba-cter-oides
bert-base-uncased	in-fer-tility	spectroscopy	ne-uro-chemist-ry	ex-ci-pie-nts	ba-cter-oides
Bio_ClinicalBERT	in-fer-tility	s-pect-ros-copy	ne-uro-chemistry	ex-ci-pie-nts	b-act-ero-ides
biobert-base-cased-v1.2	in-fer-tility	s-pect-ros-copy	ne-uro-chemistry	ex-ci-pie-nts	b-act-ero-ides
Clinical_ModernBERT	Inf-ertility	Spect-rosc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oides
ModernBERT-base	Inf-ertility	Spect-rosc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oides
distilbert-base-multilingual-cased	Inf-erti-lity	Sp-ect-ros-co-py	Neu-roch-emi-str-y	Ex-ci-pien-ts	Ba-cter-oides
mmBERT-base	Infer-tility	Spectroscopy	Neuro-chemistry	Ex-cip-ients	Bacter-oides
xlm-roberta-base	Inf-ert-ility	Spec-tros-copy	Neuro-che-mist-ry	Exc-ipi-ents	Bac-tero-ides
xlm-roberta-large	Inf-ert-ility	Spec-tros-copy	Neuro-che-mist-ry	Exc-ipi-ents	Bac-tero-ides
BioMistral-7B	In-fer-t-ility	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-acter-oid-es
gemma-3-4b-pt	In-fert-ility	Spect-roscopy	Neuro-chemistry	Ex-cip-ients	B-acter-oides
Llama3-Med42-8B	Inf-ertility	S-pect-ro-scopy	Ne-uro-chemistry	Exc-ipients	B-acter-oid-es
medgemma-4b-it	In-fert-ility	Spect-roscopy	Neuro-chemistry	Ex-cip-ients	B-acter-oides
MediPhi-Instruct	In-fer-til-ity	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-act-ero-ides
Meta-Llama-3-8B	Inf-ertility	S-pect-ro-scopy	Ne-uro-chemistry	Exc-ipients	B-acter-oid-es
Mistral-7B-Instruct-v0.1	In-fer-t-ility	Spect-ro-sc-opy	Ne-uro-chemistry	Ex-cip-ients	B-acter-oid-es
Phi-3.5-mini-instruct	In-fer-til-ity	Spect-ro-sc-opy	Ne-uro-chem-istry	Ex-cip-ients	B-act-ero-ides

Table 7: Qualitative examples of tokenization across encoder and decoder models. Medical terms are frequently fragmented into subword units that do not necessarily correspond to medically meaningful morphemes.

C Definition Retrieval - Mean Reciprocal Rank

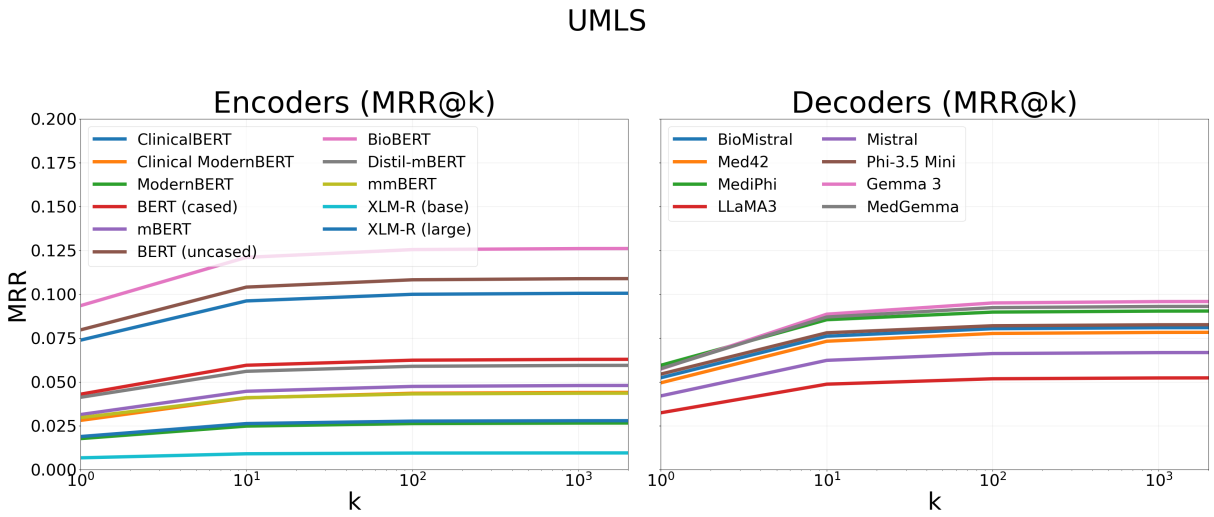


Figure 4: Definition retrieval performance on UMLS measured by MRR@k. Encoder models are shown on the left and decoder models on the right.

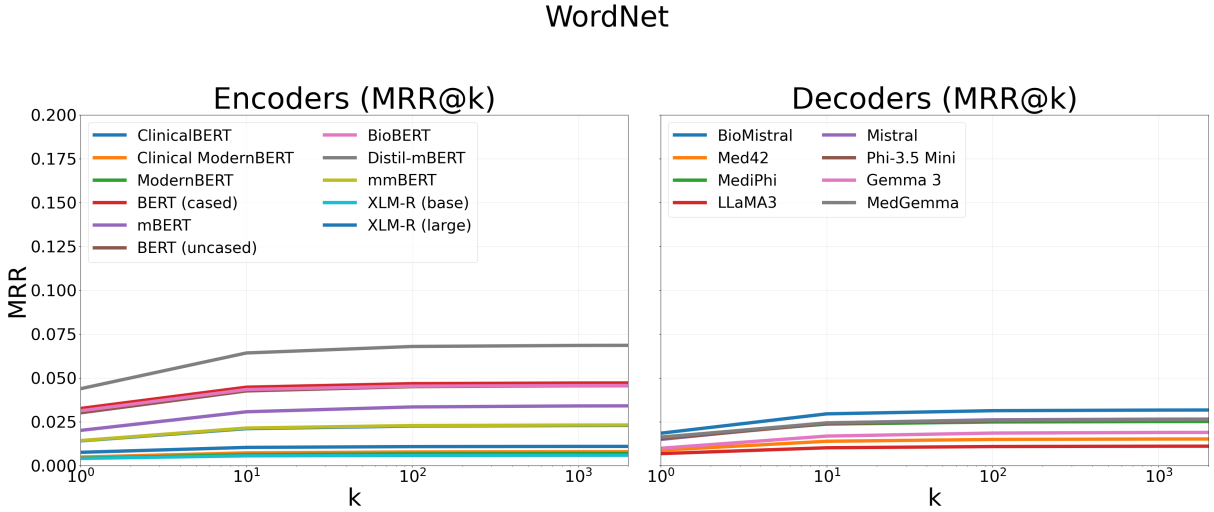


Figure 5: Definition retrieval performance on WordNet measured by MRR@k. Encoder models are shown on the left and decoder models on the right.

D Correlation Between Tokenization Granularity and Definition Retrieval

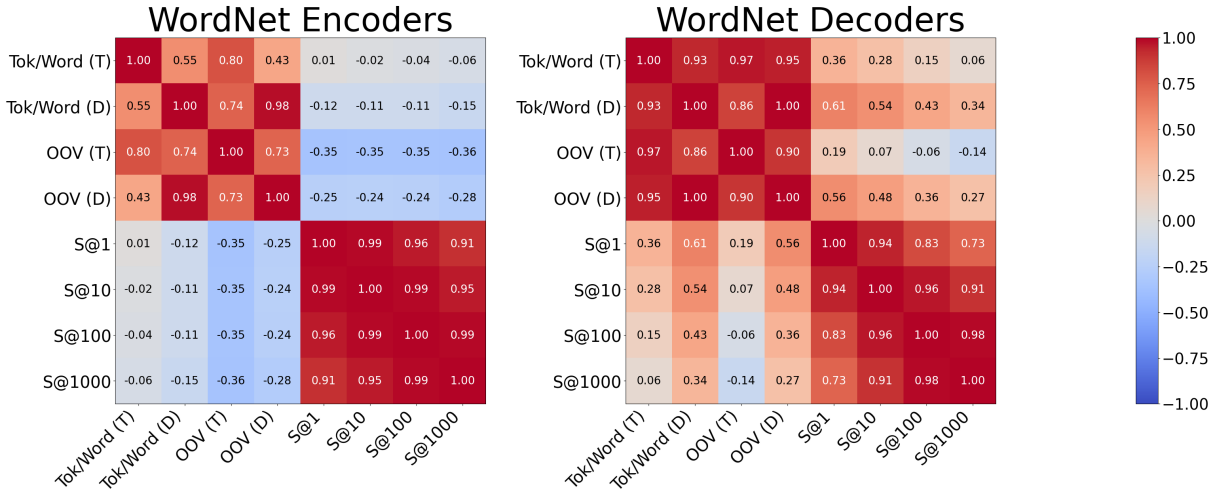


Figure 6: Pearson correlation matrices for WordNet models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (Success@k, denoted as S@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

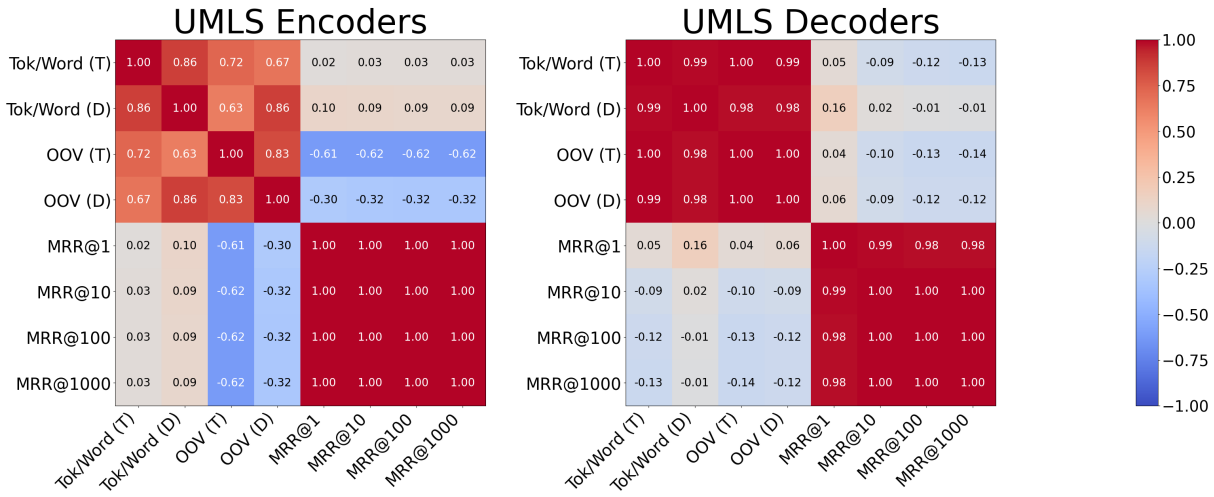


Figure 7: Pearson correlation matrices for UMLS models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (MRR@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

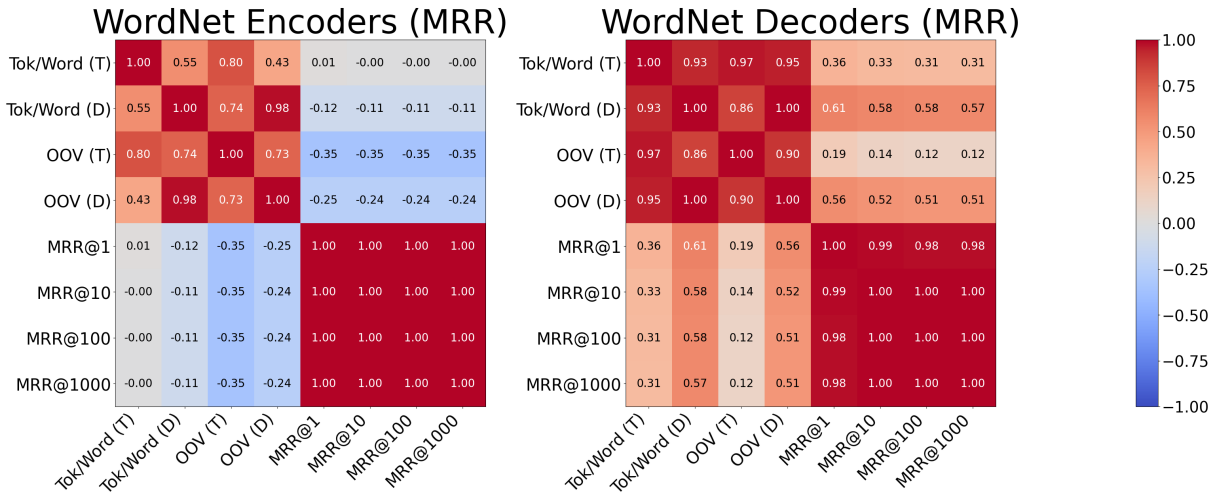


Figure 8: Pearson correlation matrices for WordNet models, separately for encoders (left) and decoders (right). The matrices report correlations between retrieval performance (MRR@k for $k \in \{1, 10, 100, 1000\}$) and tokenization granularity, including average tokens per word (Tok/Word) and out-of-vocabulary rate (OOV), computed separately for terms (T) and definitions (D).

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, Gibaeg Kim
AITRICS
harrymetsally@aitrics.com

Abstract

While Large Language Models (LLMs) have advanced clinical dialogue-to-note generation, direct transcript-only prompting can still fail in clinically important ways. In real-world physician–patient encounters, clinically salient evidence is noisy, distributed across turns, and frequently revised. Transcript-only generation is therefore vulnerable to hallucinated facts, section leakage, and final-state errors. Conversely, coarse intermediate scaffolds such as basic entity extraction fail to capture negation, temporality, and evolving problem–plan relationships.

To address this, we propose **Clinical Atomic Propositions (CAPs)**, a dialogue-aware intermediate representation for LLM-based note generation. CAPs extract standalone clinical assertions while explicitly preserving essential modifiers such as verification status, temporality, speaker source, and action type. We further introduce an optional **event consolidation layer** that reorganizes CAPs into problem-oriented care bundles, explicitly linking grounded evidence to planned actions before rendering. We evaluate sectioned-note generation as the main setting and include SOAP-template rendering as an ablation. This decomposition targets faithful, controllable note generation while producing evidence-linked structure that can support future audit and verification.

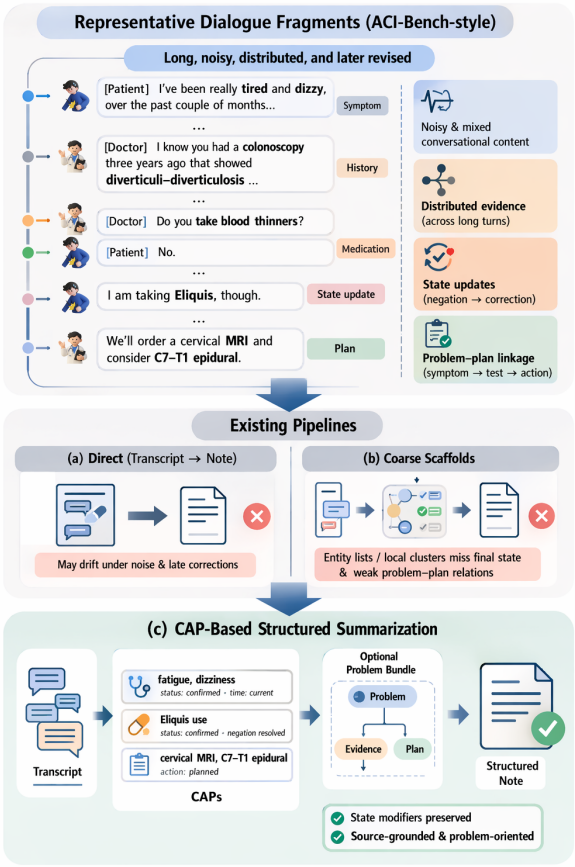


Figure 1: Why faithful dialogue-to-note generation is difficult: evidence is noisy, scattered, and later-updated; CAP makes intermediate clinical facts explicit before rendering.

1 Introduction

LLMs make it easy to prompt for a clinical note directly from a physician–patient transcript, but this convenience moves safety-critical decisions into an opaque long-context generation step (Ben Abacha et al., 2023; Giorgi et al., 2023; Krishna et al., 2021; Michalopoulos et al., 2022). The central challenge is no longer fluency alone. A useful note must be readable, *source-grounded*, clinically faithful, and section-appropriate, while preserving facts that are noisy, distributed across turns, and often revised over time. Transcript-only generation is therefore vulnerable to omission, section leakage, unsupported fill-in, and incorrect final-state tracking.

A growing line of work therefore uses *structure-first* pipelines, including section filtering, utterance clustering, and entity-centric scaffolds (Krishna et al., 2021; Michalopoulos et al., 2022; Nair et al., 2023). These approaches help by giving the final generator a smaller, more organized input. How-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 572–594
July 3-4, 2026 ©2026 Association for Computational Linguistics

ever, each scaffold can lose a different clinical relation. A proximity-based cluster may split an early statement from a later correction; an entity/status list may record that a medication is present without preserving the full claim or its timing; and a section-level filter may collect relevant text without saying which problem a finding or plan supports. As Figure 1 illustrates, the failure is often not that every relevant mention is absent, but that the *final clinically grounded state* and the problem–evidence–plan structure are not preserved.

We address this gap with **Clinical Atomic Propositions (CAPs)**, a dialogue-aware intermediate representation that converts raw dialogue into standalone, source-grounded clinical assertions. CAPs preserve the modifiers that are most critical for faithful note generation, including verification status, temporality, speaker/source, and action type. A list of correct propositions is still not necessarily a note-ready structure, so we additionally introduce an optional **event consolidation layer** that groups CAPs into problem-oriented care bundles before rendering. This separates two roles that are often conflated in transcript-only generation: preserving source-grounded facts and organizing those facts into clinically coherent assessment and plan structure.

Our work makes three main contributions:

- **Source-grounded proposition scaffold.** We introduce CAPs as a proposition-level intermediate representation for clinical dialogue-to-note generation that preserves clinically important modifiers beyond entity presence alone, including verification status, temporality, speaker/source, and action type.
- **Problem-oriented structural alignment.** We design and evaluate an optional event-consolidation step, inspired by problem-oriented documentation and lightweight FHIR semantics, that groups grounded facts into problem-centered care bundles. Our analysis characterizes when this organization helps and when compression can introduce omissions.
- **Reproducible evaluation and release.** We evaluate on ACI-Bench (Yim et al., 2023) using explicit cohort filtering criteria for reproducibility. We include prompt-based reimplementations of Cluster2Sent and MEDSUM-ENT, reuse the MEDSUM-ENT GPT-R/P/F1 metric family, and adapt the same recall/precision/F1 logic to CAP-space semantic auditing. We evaluate sectioned-note rendering as the main task and SOAP-template rendering as an ablation. We release the code, prompts, postprocessing scripts, intermediate representations, and generated sectioned/SOAP notes in a public repository: `https://github.com/sallyy1/CAP-ACL-2026`.

2 Related Work

Structure-first clinical dialogue understanding and note generation. Recent work increasingly treats doctor–patient dialogue summarization as a *structure-then-generate* problem rather than pure long-context generation (Ben Abacha et al., 2023; Giorgi et al., 2023). Cluster2Sent (Krishna et al., 2021) selects section-relevant utterances and clusters nearby evidence; MedicalSum (Michalopoulos et al., 2022) and ClinicSum (Neupane et al., 2024) use guided or retrieval-filtered SOAP inputs; and MEDSUM-ENT (Nair et al., 2023) extracts entities with present/absent/unknown status labels before summarization. Adjacent work also models clinical dialogue through task-specific labels, including note-section classification and medical-order extraction (Chen et al., 2023; Corbeil et al., 2025), demonstrating that schema-constrained outputs can facilitate clinical understanding. However, existing scaffolds are usually section-level, proximity-based, task-specific, or entity/status-based. They help organize material, but do not directly represent self-contained clinical claims with source evidence, temporal status, and problem–action linkage.

From clinical IE to proposition-level representations. Clinical information extraction traditionally maps unstructured notes into entities, modifiers, and relations, and recent LLM-based IE continues this line with clinical NER/RE models and tools such as Kiwi (Hu et al., 2026). FHIR-oriented text mining and commercial systems also show the value of interoperable extraction: for example, Azure AI Language Text Analytics for health can return a FHIR resource bundle, while Azure Health Data Services provides a managed FHIR service (Daumke et al., 2019; Microsoft, 2025, 2026). These efforts support interoperability, but a dialogue-aware note scaffold must also decide which evolving claim is finally supported, how negation and temporality apply, and how evidence/actions attach to problems. Proposition-level verification work such as VeriFact(Chung et al.,

2025) motivates this granularity by decomposing clinical text into simple statements for fact checking. CAP adapts this idea before rendering: it extracts evidence-linked, FHIR-inspired propositions from dialogue, then optionally consolidates them into problem-oriented bundles for note generation.

3 Motivation and Problem Setting

Clinical dialogue-to-note generation is difficult because clinically important evidence is noisy, distributed, and frequently revised across the encounter. This creates four recurring challenges for transcript-only or coarse-scaffold generation.

Motivation challenges.

- **Noisy dialogue:** salient evidence is interleaved with low-value conversational content.
- **Lay expression vs. over-abstraction:** patient language must be normalized without losing negation, uncertainty, temporality, or medication nuance.
- **Brittle state tracking:** key facts are revised over time, including symptom status, resolved uncertainty, medication use, and downstream plans.
- **Baseline limitations:** Direct is vulnerable to noise and section leakage, Cluster2Sent is local, and MEDSUM-ENT can miss problem-linked organization and state–plan separation.

Given a transcript X, our goal is to generate a clinical note Y that is clinically faithful, section-appropriate, and source-grounded. We ask:

Research questions.

- **RQ1: Source-grounded preservation.** Can a normalized intermediate representation better preserve clinically salient source content?
- **RQ2: Clinical state fidelity.** Can it better retain negation, uncertainty, temporality, and medication state or change?
- **RQ3: Problem-oriented organization.** Can it better support problem–evidence–plan linkage and final note structure?

Figure 2 illustrates the CAP pipeline, while Tables 1, 2, and 3 summarize the running example, FHIR-inspired CAP-type schema, and compared method scaffolds.

4 Method

4.1 Pipeline Overview

We decompose dialogue-to-note generation into a structure-first pipeline:

$$X \rightarrow C \rightarrow E \rightarrow Y, \tag{1}$$

where X is the dialogue transcript, C is a set of **Clinical Atomic Propositions (CAPs)**, E is an optional **problem-oriented consolidation plan** ("event" bundles), and Y is the rendered clinical note. CAPs provide a source-grounded content inventory, while event consolidation aligns that inventory to problem-oriented documentation practice (POMR/SOAP) by grouping evidence and actions under an anchor problem. The renderer then writes the final sectioned or SOAP note from the selected scaffold.

4.2 Clinical Atomic Propositions

A **Clinical Atomic Proposition (CAP)** is a standalone clinical assertion that is verifiable against the dialogue and close to the semantic grain of a note sentence. A CAP is not just an entity mention. It records the clinical claim plus state modifiers such as verification status, temporality, speaker/source, numeric attributes, and whether the claim describes current state or an intended action. CAPs are extracted automatically by prompted, schema-constrained JSON generation; we do not manually author them at test time.

The CAP schema is *FHIR-inspired* rather than FHIR-native (Health Level Seven International, 2019; Ayaz et al., 2021). Table 2 lists the canonical `cap_types` used in this paper. While FHIR is designed primarily for interoperable health-data exchange, CAP adopts its resource-oriented separation of clinical concepts to define a compact middle schema for note generation. This schema separates condition-like problems, observation-like evidence, current medication state, medication changes, and planned actions. The distinction is central for dialogue: for example, "taking Eliquis" and "order an MRI" should not be collapsed into the same kind of fact. The small CAP-extraction dev check is reported in Appendix B.

4.3 Event Consolidation and Rendering

Consistent with the Problem-Oriented Medical Record (POMR) (Weed, 1968), clinical notes are typically organized around problems with linked

Stage	D2N026-style late medication update
Dialogue X	"Do you take blood thinners?" → "No." → later correction: "I am taking **Eliquis**, though." The physician also discusses a **cervical MRI** and tentative **C7–T1 epidural**.
CAPs C	Separate source-grounded propositions preserve the corrected final state, e.g., `MedicationStatement: patient currently taking Eliquis; Order: cervical MRI planned; ProcedurePlan: C7-T1 epidural discussed`.
Bundle E	The bundler groups these CAPs under an anchor problem such as `Problem: cervical radiculopathy`, with linked evidence and plan slots.
Note Y	The renderer realizes the corrected state and plan in prose, e.g., Eliquis use, cervical MRI, and possible C7–T1 epidural.

Table 1: Running example for the proposed $X \rightarrow C \rightarrow E \rightarrow Y$ pipeline.

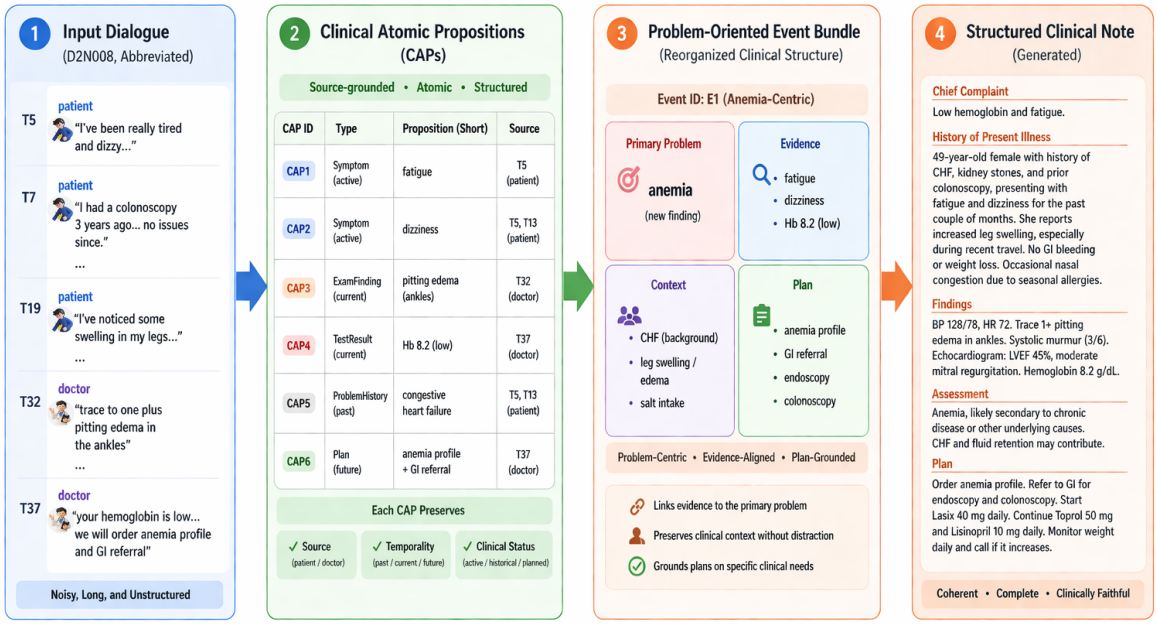


Figure 2: Overview of the CAP-based pipeline. Dialogue is first converted into source-grounded atomic propositions (CAPs), which are then reorganized into a problem-oriented event structure. This structured representation enables coherent and clinically faithful note generation.

evidence and plans, a structure commonly operationalized through the SOAP (Buchanan, 2017) template. Our optional event layer turns a flat CAP list into **problem-oriented bundles**: each bundle anchors on a problem and links related subjective evidence, objective evidence, and plan actions. We treat this as an alignment module rather than a guaranteed quality booster. It can make problem–plan linkage explicit, but it can also over-compress or omit facts if the bundling is too aggressive.

The final renderer receives the chosen scaffold and writes a sectioned or SOAP note. In the main CAP and CAP+Event settings, the transcript is also available for surface realization; transcript-free ablations in Appendix E test how much the scaffold can support rendering on its own. Prompt templates and baseline reimplementation details are provided in Appendix C.

5 Experiments

Dataset. We use ACI-Bench (Yim et al., 2023), a publicly documented ambient clinical intelligence benchmark for automatic visit-note generation from doctor–patient conversations. The original paper describes ACI-Bench as a *shared open dataset* that is *freely available*; we follow that release as our primary data source. Following the benchmark setup, we use the provided conversation transcripts as model input and the provided reference notes as evaluation targets.

To avoid ambiguity about supervision, we do not create new gold notes or re-annotate the benchmark. Reference-note provenance and annotation protocol are therefore inherited from ACI-Bench and are documented in Yim et al. (2023). For reproducibility, we start from 207 benchmark encounters

CAP type	Intended semantics	FHIR-inspired analogue	Key state fields
`ChiefComplaint`	Reason for visit / presenting concern	Encounter reason (conceptual)	`verification_status`, `temporality`
`Problem`	Active problem / diagnosis / clinical concern	Condition-like	`verification_status`, `clinical_status`, `temporality`
`ProblemHistory`	Historical problem context	Condition-like (historical)	`clinical_status=historical`
`ExamFinding`	Physical exam finding / symptom-style observation	Observation-like	`verification_status` (e.g., negated), `temporality`
`TestResult`	Lab / imaging result	Observation/Report-like	`verification_status`, numeric `attributes`
`MedicationStatement`	Current medication state (what the patient is taking)	MedicationStatement-like	`clinical_status=active` vs. historical
`MedicationRequest`	Intended medication start/stop/change	MedicationRequest-like	`clinical_status=planned`
`Order`	Concrete order (labs, imaging, referral)	ServiceRequest-like	`clinical_status=planned`
`FollowUp`	Follow-up timing / appointment plan	CarePlan/Appointment-like	`temporality`, `clinical_status=planned`
`Counseling`	Advice / behavior counseling / education	CarePlan-activity-like	`temporality`, `clinical_status`

Table 2: Canonical `cap_types` and intended semantics. The mapping is a lightweight FHIR-inspired abstraction, not a full FHIR implementation. CAPs additionally store `proposition_text`, `canonical_concept`, `evidence`, and optional `linked_problem` pointers used by event consolidation.

Method	Input to downstream renderer	Intermediate scaffold	Unit	Value / limitation
Direct (base)	Transcript X	None	None	+ fluent baseline – noise, leakage, state drift
MEDSUM-ENT (prior)	Transcript X + entity/status scaffold	6-category plan (*Demographics/SDOH*, *Intent*, *Positives*, *Negatives*, *Unknowns*, *History*) $\times$ {*present*, *absent*, *unknown*}	Entity/status	+ status cues – weak problem–plan linkage
Cluster2Sent (prior)	Section-targeted evidence units only (no X)	Utterance clusters (proximity; τ=2)	Utterance cluster	+ local coherence – long-range/state ambiguity
CAP (ours)	Transcript X + CAP set C	Atomic propositions with clinical modifiers	Proposition	+ grounded state fidelity – needs rendering for organization
CAP+Event (ours)	Transcript X + C + bundle plan E	Problem-oriented bundles linking evidence and actions	Problem bundle	+ problem–plan organization – compression can omit

Table 3: Strategic positioning of the five main compared methods and their intermediate scaffolds. Method-label shading matches Table 6: gray marks the Direct baseline, blue marks transcript+entity/proposition scaffold methods, and beige marks clustering/bundling scaffold methods. Transcript-free ablations are reported in Appendix E.

and apply a predefined CAP-quality cohort filter, yielding a final 197-case cohort (Appendix A lists excluded IDs and exact criteria). Our release package focuses on reproducible prompts, intermediate representations, and scoring scripts for this fixed cohort.

Setup and methods. We evaluate on the final 197-case ACI-Bench cohort after CAP extraction and predefined quality filtering; cohort details and excluded IDs are in Appendix A. The main experiments use a sectioned-note template, with SOAP ablations in Appendix D. We compare

Aspect	GPT-R/P/F1 (Nair-style)	semCAP-R/P/F1 (ours)
Primary purpose	Reference-oriented overlap check	Source-grounded proposition audit
Compared units	Medical concepts with status cues; extracted under MEDSUM-ENT-style categories (*Demographics/SDOH, Intent, Positives, Negatives, Unknowns, History*)	Proposition-level CAP units aligned to dialogue evidence
Source of truth	Reference note (Generated note ↔ Reference note)	Dialogue-grounded CAP/audited propositions (Dialogue transcript ↔ Generated note)
Rewards	Coverage of reference-like concepts and *Present/Absent/Unknown* status consistency	Preservation of grounded assertions and key modifiers (*verification, temporality, source, action type*, etc.)
May miss	Direct dialogue grounding and problem–plan linkage even when concept overlap is high	Broader narrative/note-quality aspects
Role in this paper	Complementary note-level metric	Primary targeted metric for CAP claim
Limitation	Reference-oriented, not grounding-specific	May be partially aligned with CAP representation

Table 4: Comparison of Nair-style GPT metrics and semCAP. We report both because they capture complementary aspects: reference-note overlap vs. dialogue-grounded faithfulness.

View (D2N026)	Concrete extracted units (abridged)	Count / score
Transcript CAP units (semCAP-R)	"Left arm/hand pain for about two weeks"; "Hand weakness and grip difficulty"; "Patient is taking Eliquis"; "Clinician orders cervical MRI", ...	25 CAPs
Generated-note CAP units (semCAP-P, CAP note)	"The patient reports left arm pain."; "...hand pain began approximately two weeks ago."; "I will order a cervical MRI."; "...tentatively schedule a left C7–T1 epidural.", ...	21 CAPs
Reference-note concepts (GPT-R)	*left arm pain*, *gabapentin*, *cervical MRI*, *left C7–T1 epidural*, *blood thinners*, ...	45 concepts
Generated-note concepts (GPT-P, CAP note)	*left arm pain*, *hand weakness*, *cervical radiculopathy*, *cervical MRI*, *C7–T1 epidural*, ...	24 concepts
Case-level metric snapshot (sectioned template)		
Direct	Nair GPT-F1 / semCAP-F1	0.098 / 0.000
CAP	Nair GPT-F1 / semCAP-F1	0.488 / 0.543

Table 5: Concrete example of the two metric families on D2N026 (values from released case-level outputs). Nair-style GPT metrics compare reference-note vs generated-note concepts, while semCAP compares transcript-grounded vs generated-note CAP propositions. The Direct semCAP-F1 value of 0.000 is a case-level boundary example under strict proposition matching, not a global trend. Full R/P/F1 values are reported in Table 6.

Direct, **Cluster2Sent** (Krishna et al., 2021), **MEDSUM-ENT** (Nair et al., 2023), **CAP**, and **CAP+Event**; Table 3 summarizes their inputs, scaffolds, and limitations. All methods use the same downstream generator, and all CAP extraction/rendering stages use the open-weight **Gemma-3-4B-Instruct** model (Hugging Face ID: `ISTA-DASLab/gemma-3-4b-it-GPTQ-4b-128g`) to reflect privacy-constrained clinical deployment. Baseline details and prompts are in Appendix C; transcript-free ablations and cost analyses are in Appendices E and F.

Metrics. We separate two automatic metric families. **GPT-R/P/F1 (Nair)** follows the MEDSUM-ENT GPT-style concept matching framework (Nair et al., 2023). Here, a *concept* denotes a medically relevant item (e.g., symptom, condition, medication, test/procedure, status cue). The framework uses a two-stage LLM process: a concept extractor first identifies candidate concepts in generated/reference notes, then a verifier judges concept matches across notes under relaxed semantic equivalence, explicitly allowing clinically valid rephrasings. GPT-recall measures reference concepts recovered by the generated note, GPT-precision measures generated concepts supported by the reference note, and GPT-F1 is their harmonic mean. This family is therefore *reference-oriented* and primarily reflects concept-level note overlap.

semCAP-R/P/F1 (ours) is our proposition-grounded semantic audit, measuring whether transcript-derived CAP propositions are preserved

Method	semCAP-R (ours)↑	semCAP-P (ours)↑	semCAP-F1 (ours)↑	GPT-R (Nair)↑	GPT-P (Nair)↑	GPT-F1 (Nair)↑
Direct (base)	0.574	0.175	0.211	0.638	0.808	**0.703**
MEDSUM-ENT (prior)	0.589	0.168	0.205	**0.640**	0.761	0.687
Cluster2Sent (prior)	0.522	0.178	0.204	0.587	**0.813**	0.672
CAP (ours)	**0.694**	**0.212**	**0.256**	0.632	0.795	0.695
CAP+Event (ours)	0.570	0.181	0.214	0.578	0.773	0.649

Table 6: Results on the 197-case ACI-Bench cohort. (Sectioned-template) Method-label shading groups the Direct baseline (gray), transcript+entity/proposition scaffold methods (blue), and clustering/bundling scaffold methods (beige). semCAP metrics are our source-grounded proposition audit; GPT metrics follow the MEDSUM-ENT concept-matching family (Nair et al., 2023). Broader checklist and PDSQI scores are reported in Appendix D.

Method vs. Direct	**semCAP-R** W/T/L, Δ	**semCAP-F1** W/T/L, Δ	**GPT-F1 (Nair)** W/T/L, Δ
MEDSUM-ENT	111/13/73, +0.015	55/99/43, -0.006	101/4/92, -0.016
Cluster2Sent	60/15/122, -0.052	47/97/53, -0.007	64/2/131, -0.031
CAP (ours)	**170/4/23, +0.120**	**83/97/17, +0.046**	85/2/110, -0.008
CAP+Event (ours)	103/15/79, -0.004	53/97/47, +0.003	83/2/112, -0.054

Table 7: Case-level win/tie/loss against Direct. (Sectioned-template) Full counts corresponding to Figure 3.

and supported in the generated note, including clinically critical modifiers (negation/state, temporality, and source grounding). semCAP is thus targeted to source-grounded faithfulness rather than broad note-quality similarity. Because semCAP is aligned with our representation, some metric-method alignment bias may remain; we therefore report semCAP jointly with the reference-oriented GPT metrics and interpret them as complementary, not interchangeable, signals. All automated metric evaluation uses **OpenAI GPT-5.1** (API model ID: `gpt-5.1-2025-11-13`) as judge. Additional checklist/PDSQI analyses are in Appendix D; Tables 4–5 summarize metric differences and one case-level example.

Main results on the 197-case cohort. Table 6 shows that CAP is strongest on all source-grounded semCAP metrics, improving semCAP-R from 0.574 for Direct to 0.694 (+0.120 absolute) and semCAP-F1 from 0.211 to 0.256. On the MEDSUM-ENT GPT concept metrics, CAP remains close to Direct (GPT-F1 0.695 vs. 0.703), while MEDSUM-ENT narrowly leads on GPT-R and Cluster2Sent on GPT-P. We therefore do not claim uniform superiority on generic note/reference overlap. Instead, the main quantitative finding is more specific: CAP improves preservation of transcript-grounded clinical propositions while maintaining competitive concept-level note similarity.

Figure 3 further clarifies the aggregate pattern. CAP wins over Direct on semCAP-R in 86.3% of cases and loses in only 11.7%, while its GPT-F1 is nearly tied on average. This supports our intended positioning: CAP is not primarily a fluency or broad quality optimizer, but a source-grounded content scaffold for preserving clinically salient facts. CAP+Event does not uniformly outperform CAP, reinforcing our view that event consolidation is an alignment module whose benefit is case dependent. Detailed win/tie/loss counts for semCAP-R/F1 and GPT-F1 are reported in Table 7.

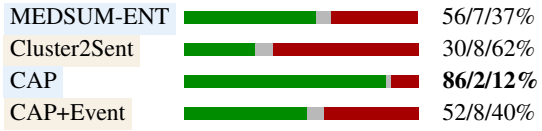


Figure 3: Case-level semCAP-R win/tie/loss rates against Direct. (Sectioned-template) Label shading matches Table 6: blue marks transcript+scaffold methods and beige marks clustering/bundling methods. Green/gray/red bars indicate wins/ties/losses.

Category-level GPT submetrics were also inspected during development, but they were sparse and unstable on this cohort. We therefore report them only as internal diagnostics and keep the main claims focused on semCAP and GPT aggregate metrics.

Qualitative evidence. D2N008 is our main qualitative case because it mixes a new anemia workup with background CHF, edema, prior GI history,

Method	Representative output snippet
Direct	Clinically plausible but structurally shallow: anemia is identified and the GI workup is mentioned, yet the CHF–anemia relation remains vague and the note never fully reconstructs anemia as the dominant organizing problem.
Cluster2Sent	Preserves fatigue and dizziness, but also keeps lower-priority content such as [incorrect emphasis] "mild nasal congestion related to seasonal allergies", producing *salience dilution* rather than a clear anemia-centered hierarchy.
MEDSUM-ENT	Entity-centric planning captures concept/status cues but does not explicitly reconstruct anemia-centered problem–evidence–plan structure; in this case the sectioned renderer can under-generate to an *empty skeleton* (see Appendix H).
CAP	Stronger structure and symptom/history coverage, but it introduces [incorrect attribution] "CHF, diverticulosis, and seasonal allergies are contributing factors", which over-expands peripheral findings into the anemia assessment.
CAP+Event (ours)	Re-centers the note on anemia as the primary problem and best preserves problem–plan linkage, including [correct] "ordering an anemia profile" and [correct] "refer ... for an endoscopy and repeat colonoscopy".

Table 8: Main qualitative case study on D2N008. Under a multi-condition scenario with background CHF and a new anemia workup, CAP-based structured generation better preserves clinically relevant problem–evidence–plan linkage than transcript-only, utterance-clustering, and entity-centric baselines (more details in Appendix H).

Case	Scenario	Observed CAP+Event behavior	ΔsemCAP-R	ΔGPT-F1 (Nair)
D2N207	Multi-condition/cardiorenal visit with CKD context and possible gout flare	Produces two bundles (*chronic kidney disease*; *possible gout flare*), improving organization and grounded recall over CAP.	$+0.218$	$+0.235$
D2N153	Localized musculoskeletal injury follow-up	Collapses to one narrow *toe injury* bundle and under-generates an empty sectioned skeleton, losing much of CAP's preserved content.	-0.714	-0.771

Table 9: Representative boundary cases for event consolidation (CAP+Event minus CAP). We report one source-grounded metric (semCAP-R) and one MEDSUM-ENT-style concept metric (GPT-F1) to show the preservation/compression trade-off under two complementary metric families.

and low-value allergy content. The key challenge is whether the note reconstructs anemia as the organizing problem and keeps the anemia profile and GI referral linked to that problem.

Table 8 illustrates the pattern. Direct remains plausible but weakly organized; Cluster2Sent preserves local evidence but over-emphasizes peripheral allergy content; MEDSUM-ENT captures entity/status cues but does not explicitly reconstruct anemia-centered problem–evidence–plan structure. CAP provides a better source-grounded evidence inventory, and CAP+Event most clearly links the anemia problem to the anemia profile and GI workup. Full qualitative cases for D2N008, D2N026, and D2N024 are shown in Appendices H, I, and J, respectively; D2N026 highlights late medication-state correction, while D2N024 is a mixed boundary case.

When does event consolidation help? Event consolidation is case dependent. In aggregate, CAP+Event underperforms CAP on semCAP-R, suggesting that bundling can introduce omission through compression. However, stratifying by the number of bundles produced by the consolidator reveals a more nuanced pattern: when the consolidator produces a modest multi-bundle plan (E with two bundles), organization-oriented scores improve slightly on average (mean ΔPDSQI $+0.04$; mean ΔF-CL(A+B+C) $+0.03$), whereas single-bundle outputs show consistent degradation across grounding and quality metrics (Appendix G). Table 9 shows this boundary behavior in representative cases: consolidation helps when distinct threads are preserved, but can fail when content is over-collapsed into a narrow anchor.

Clinician deep review. Two clinicians with medical training and experience reviewing clinical notes independently performed a focused deep review on three cases representing clear gain, mixed trade-off, and boundary behavior. Together with the reviewers, we designed a checklist aligned with our Motivation challenges and Research Questions to directly probe three target axes: (A) content preservation, (B) clinical state fidelity, and (C)

problem-oriented organization. Reviewers rated eight items on a 1–5 Likert scale (1=poor, 5=excellent) and flagged clinically meaningful omission/hallucination/major errors (yes/no). An optional overall usability rating was collected for context but excluded from our main faithfulness aggregate. Full rubric details, scores, and anonymized comments are in Appendix K (Figure 5).

6 Discussion

Our framework separates two sources of difficulty in clinical note generation: *factual recovery* and *document organization*. CAPs preserve clinically meaningful atomic facts from noisy dialogue, while event consolidation reorganizes grounded facts into problem-level bundles. This separation improves controllability and helps localize failures to extraction, consolidation, or rendering. It also clarifies why semCAP and GPT metrics need not move identically: the former audits transcript-grounded proposition fidelity, whereas the latter measures reference-note concept overlap.

CAP extraction also exposes a trade-off that is often hidden in transcript-only generation: *what to preserve* versus *what to suppress*. More aggressive filtering reduces small talk and background chatter, but can under-extract evidence needed later for bundling and planning. Our extractor development results show this precision–recall tension while preserving key modifiers well on matched propositions, especially verification status (Appendix B). Future work should add explicit salience/priority modeling, including severity and clinical consequentiality, so noise suppression does not become brittle omission.

More broadly, fluency alone is not sufficient: a natural-sounding note can still be clinically misleading. CAPs provide an evidence-linked structure that may also support real-time clinician correction or post-generation audit/verification, though we leave those uses to future work. Remaining CAP+Event errors show that faithful rendering from structured clinical scaffolds is still an open biomedical NLP problem, particularly when organization requires compressing heterogeneous dialogue evidence without losing clinically relevant detail.

7 Conclusion

We presented CAP, a structure-first framework for clinical dialogue-to-note generation. CAP introduces a source-grounded proposition layer that preserves clinically meaningful assertions from noisy dialogue, and an optional event consolidation layer that organizes these assertions into problem-oriented care bundles. Together, these components provide a more controlled scaffold for note generation than transcript-only or coarse structured approaches, with the clearest gains appearing in source-grounded faithfulness, clinical state preservation, and problem-oriented note construction.

Limitations

Our work has several limitations. semCAP is a targeted, representation-aligned audit metric rather than a standalone measure of overall clinical note quality, and our experiments are limited to a single benchmark, leaving broader generalization across specialties, note styles, and clinical settings for future work. The event layer is inspired by clinical documentation practice and lightweight FHIR semantics (Health Level Seven International, 2019; Ayaz et al., 2021), but is not a formal interoperability standard. Practically, the multi-stage prompted pipeline can propagate errors from CAP extraction through event clustering to rendering, and it adds non-trivial latency and token overhead relative to Direct. Finally, the renderer can still over-abstract, introduce unsupported content, or under-generate an *empty skeleton* note in a small fraction of cases; we report these failures rather than filtering them out, and a key next step is human-in-the-loop correction at the CAP stage before final rendering.

References

Muhammad Ayaz, Muhammad F. Pasha, Mohammed Y. Alzahrani, Rahmat Budiarto, and Deris Stiawan. 2021. The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7):e21929.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302.

Joel Buchanan. 2017. Accelerating the benefits of the problem oriented medical record. *Applied Clinical Informatics*, 8(1):180–190.

Zhuohao Chen, Jangwon Kim, Yang Liu, and Shrikanth Narayanan. 2023. Clinical note section classification

on doctor-patient conversations in low-resourced settings. In *Proceedings of the Third Workshop on NLP for Medical Conversations*, pages 1–12.

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, and 10 others. 2025. VeriFact: Verifying facts in LLM-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Jean-Philippe Corbeil, Asma Ben Abacha, Jerome Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and Francois Beaulieu. 2025. Overview of the MEDIQA-OE 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 11–16.

Philipp Daumke, Kai U. Heitmann, Simone Heckmann, Catalina Martínez-Costa, and Stefan Schulz. 2019. Clinical text mining on FHIR. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, volume 264, pages 83–87.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. WangLab at MEDIQA-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334.

Health Level Seven International. 2019. FHIR release 4 (R4). `https://www.hl7.org/fhir/r4/`. Accessed 2026-04-16.

Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K. Keloth, Vincent J. Zhang, Ruey-Ling Weng, Cathy Shyr, Qingyu Chen, Xiaoqian Jiang, Kirk E. Roberts, and Hua Xu. 2026. Information extraction from clinical notes: Are we ready to switch to large language models? *Journal of the American Medical Informatics Association*, 33(3):553–562.

Kundan Krishna, Yixuan Zhang, Byron C. Wallace, and Mohit Iyyer. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972.

George Michalopoulos, Kyle B. Mahowald, Thomas Lin, Hyo-Jung Lee, Ke Wang, Steven Bethard, and Meliha Yetisgen. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Microsoft. 2025. Utilizing fast healthcare interoperability resources (FHIR) structuring in text analytics for health. `https://learn.microsoft.com/en-us/azure/ai-services/language-service/text-analytics-for-health/concepts/fhir`. Last updated 2025-12-05; Accessed 2026-05-19.

Microsoft. 2026. What is the FHIR service in azure health data services? `https://learn.microsoft.com/en-us/azure/healthcare-apis/fhir/overview`. Last updated 2026-05-04; Accessed 2026-05-19.

Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217.

Subash Neupane, Himanshu Tripathi, Shaswata Mitra, Sean Bozorgzad, Sudip Mittal, Shahram Rahimi, and Amin Amirlatifi. 2024. ClinicSum: Utilizing language models for generating clinical summaries from patient-doctor conversations. In *Proceedings of the 2024 IEEE International Conference on Big Data*, pages 5050–5059.

Lawrence L. Weed. 1968. Medical records that guide and teach. *New England Journal of Medicine*, 278(12):652–657.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *arXiv preprint arXiv:2306.02022*.

A Cohort Selection and Evaluation Protocol

Implementation details. All CAP extraction and note rendering stages used the same open-weight generator, Gemma-3-4B-Instruct (Hugging Face model ID: `ISTA-DASLab/gemma-3-4b-it-GPTQ-4b-128g`). We served the model with tensor parallel size 1, maximum context length 32,768, GPU memory utilization 0.9, and prefix caching enabled. In our runs, inference was executed on a single NVIDIA RTX Ada 6000 or RTX A6000 GPU. We set `max-num-seqs` to 32 and used a local Hugging Face cache under the workspace directory.

Evaluation model. All automated metric evaluation used **OpenAI GPT-5.1** (API model ID: `gpt-5.1-2025-11-13`) for concept-level matching, semantic CAP auditing, and checklist scoring, while the generator was fixed across methods; these API calls were used only for offline evaluation on de-identified benchmark text, not for deployment-time generation.

We began from 207 ACI-Bench encounters and removed four parse-failure cases from stage-C extraction (D2N022, D2N059, D2N088, D2N156), leaving 203 cases. We then applied a predefined CAP-quality filter and excluded six additional cases with `transcript_cap_count`=0 or `reference_cap_count`≤ 1 (D2N010, D2N053, D2N085, D2N112, D2N126, D2N187), yielding the final 197-case cohort for aggregate metrics; D2N010 is retained only as a qualitative boundary example. Two SOAP generations were empty (D2N031 for MEDSUM-ENT; D2N181 for Cluster2Sent), and we keep them as observed system behavior rather than filtering them out.

Empty structured outputs. Because the renderer must return a structured JSON object for SOAP/sectioned templates and may emit empty strings for unsupported sections, a failure mode is an *empty skeleton* note (all keys present, section content empty), which typically appears when long or heterogeneous scaffolds are hard to map under strict format constraints. On the final 197-case cohort, this occurred most frequently for MEDSUM-ENT (9/197 in sectioned; 12/197 in SOAP; 17 unique cases across templates) and more rarely for other method/template pairs ($\leq 5/197$ each). We treat these as observed generation failures and keep them in evaluation rather than filtering them out.

B CAP Extraction Development Benchmark

To reduce the risk of silently propagating extraction errors through the modular pipeline, we performed a small internal dev evaluation focused on stage C (CAP extraction) in $X \rightarrow C \rightarrow E \rightarrow Y$. We used two clinician-labeled seed CAP sets (D2N008, D2N026) and a small derived "silver" set constructed under the same rubric for prompt-iteration diagnostics (not final benchmark claims). We score extraction with relaxed semantic concept matching (threshold 0.35), reporting concept-level precision/recall/F1 and matched-pair modifier accuracy. Table 10 summarizes the final extractor used in this paper and the core precision–recall trade-off.

C Prompt Templates and Baseline Reimplementation

To facilitate reproducibility, we summarize the prompt templates and intermediate scaffold designs used for the main compared methods. Our implementation follows the structure-first intent of prior work while keeping the downstream renderer fixed so that differences primarily reflect scaffold quality.

Cluster2Sent-inspired scaffold. Following Krishna et al. (2021), we (i) select section-targeted evidence units from the dialogue, (ii) deterministically cluster nearby evidence by turn proximity (we use τ=2), and (iii) prompt the renderer to generate one concise sentence (or bullet) per cluster and place it in the appropriate note section. The crucial constraint is that the renderer is asked to use only the extracted evidence units as source material, which reduces transcript noise but can miss long-range dependencies.

```
[Section: History of Present Illness] [Cluster 1]
Left arm/hand pain started about two weeks ago;
hand weakness and grip difficulty.
[Section: Findings] [Cluster 2] Cervical source
is suspected based on exam context.
[Section: Plan] [Cluster 3] Order cervical MRI;
tentative left C7-T1 epidural.
[Section: Medications] [Cluster 4] Later turn:
patient reports taking Eliquis.
```

MEDSUM-ENT-inspired scaffold. Following Nair et al. (2023), we extract clinically relevant concepts with status cues (*present*/*absent*/*unknown*) and organize them into six planning categories (*Demographics/SDOH*, *Patient Intent*, *Pertinent Positives*, *Pertinent Negatives*, *Pertinent Unknowns*, *Medical History*). We apply an *unknown resolver*

Dev set	n	Pred	Gold	P	R	F1	VerifAcc	ClinStateAcc	TempAcc	Unsupported
Gold seeds (clinicians)	2	24.0	35.0	0.546	0.369	0.439	0.888	0.843	0.685	11.0
Silver dev (derived)	19	24.4	48.6	0.425	0.207	0.268	0.659	0.837	0.769	14.2

Table 10: CAP extraction dev benchmark for stage C. Pred/Gold/Unsupported are per-case averages. P/R/F1 are concept-level match scores under relaxed semantic matching (threshold 0.35). Modifier accuracies (verification status, clinical status, temporality) are computed on matched CAP pairs. The derived silver set is used only for prompt-iteration diagnostics.

that removes an earlier *unknown* mention when the same concept is later resolved as *present* or *absent*. This reimplementation preserves the original multi-stage spirit of MEDSUM-ENT: prompted structured extraction, lightweight normalization/postprocessing, and a separate rendering step.

```
[Category: Pertinent Positives] [Status: Present]
The patient reports fatigue for the past two months.
[Category: Pertinent Negatives] [Status: Absent]
The patient denies blood in stool.
[Category: Pertinent Unknowns] [Status: Unknown]
The patient is unsure about recent weight change.
```

MEDSUM-ENT GPT metrics and our semCAP adaptation. For note-level automated evaluation, we follow Nair et al. (2023) in using a GPT-based matching framework that computes **GPT-recall**, **GPT-precision**, and **GPT-F1**. Operationally, this framework uses two LLM roles: (i) an extractor that identifies medically relevant *concepts* from each note (generated/reference), where concepts are item-level clinical units (e.g., symptoms, diagnoses, medications, tests/procedures, and associated status cues), and (ii) a verifier that determines cross-note matches under relaxed semantic equivalence. This design allows clinically equivalent rephrasings beyond strict string overlap. GPT-R is the fraction of reference concepts recovered by the generated note, GPT-P is the fraction of generated concepts supported by the reference, and GPT-F1 summarizes both.

Our contribution is to adapt this evaluation logic from entity/category space into proposition space. Concretely, we replace concept-level units with CAP propositions and score whether transcript-grounded propositions are preserved in the generated note. We therefore report **semCAP-R**, **semCAP-P**, and **semCAP-F1** as CAP-space counterparts. The semantic audit further allows partial support and contradiction labels, making it more sensitive to source-grounded faithfulness, final-state errors, and state-update failures than pure concept overlap. At the same time, semCAP is representation-aligned by design, so we report it together with Nair-style GPT metrics as a complementary pair: semCAP for dialogue-grounded proposition preservation, GPT metrics for reference-note concept overlap.

CAP extraction and event planning. CAP extraction uses a dialogue-aware JSON schema that enforces atomic, verifiable propositions and preserves clinically important modifiers (negation, temporality, laterality, numeric values, severity, and plan intent). Event planning groups CAPs into problem-oriented bundles with linked evidence and actions, designed to improve renderability while keeping source grounding explicit.

```
CAP (Problem): The patient reports left hand weakness for two weeks.
CAP (MedicationStatement): The patient is taking Eliquis.
CAP (Order): The clinician orders a cervical MRI.
CAP (Plan): The clinician tentatively plans a left C7–T1 epidural.
```

Code and prompts. The full implementation, including prompts for scaffold extraction, rendering, and evaluation, is publicly available together with the postprocessing rules and scripts needed to reproduce our results. We also release intermediate artifacts—extracted scaffolds, CAPs, event plans, and generated sectioned/SOAP notes—for the final filtered cohort. The code, prompts, and reproducibility artifacts for this paper are available at https://github.com/sallyy1/CAP-ACL-2026.

Pipeline	Avg latency (s)	Rel. latency	Avg input toks	Avg output toks	Avg total toks	Rel. total toks
Direct ($X \to Y$)	10.03	1.00	2988.03	808.83	3796.87	1.00
CAP ($X \to C \to Y$)	60.71	6.06	20799.83	4255.33	25055.17	6.60
CAP+Event ($X \to C \to E \to Y$)	64.50	6.43	20634.17	4445.33	25079.50	6.61

Table 13: Deployment-path latency/token proxy on 30 cases; relative values are normalized to Direct.

D Ablation: SOAP-template Rendering

This section complements the main sectioned-template results with SOAP-template diagnostics. Table 11 reports SOAP-template aggregate metrics for the same five methods as the main experiment. Figure 4 then visualizes case-level SOAP semCAP-R win/tie/loss rates against Direct. Sectioned-template full win/tie/loss counts are reported in the main text (Table 7).

Method	semCAP-R (ours)↑	semCAP-P (ours)↑	semCAP-F1 (ours)↑	GPT-R (Nair)↑	GPT-P (Nair)↑	GPT-F1 (Nair)↑
Direct (base)	0.580	0.170	0.203	**0.636**	0.785	**0.693**
MEDSUM-ENT (prior)	0.546	0.158	0.191	0.596	0.737	0.649
Cluster2Sent (prior)	0.470	0.176	0.192	0.527	**0.791**	0.617
CAP (ours)	**0.660**	**0.210**	**0.246**	0.548	0.757	0.626
CAP+Event (ours)	0.520	0.184	0.196	0.508	0.771	0.600

Table 11: Results on the 197-case ACI-Bench cohort. (SOAP-template).

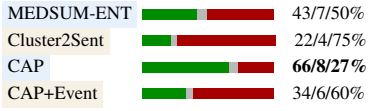


Figure 4: Case-level semCAP-R win/tie/loss rates against Direct. (SOAP-template).

E Ablation: Transcript-Free Rendering

To isolate how much of note generation can be supported by the intermediate representation alone, we evaluate two transcript-free ablations: **CAP-only** (renderer receives only the CAP set C) and **CAP+Event-only** (renderer receives C plus the bundle plan E), both without transcript X. These are not intended as competitive baselines; rather, they stress-test the *renderability* of intermediate scaffolds under a strict source constraint. Table 12 summarizes the sectioned-template means for these transcript-free conditions.

Method	semCAP-R (ours)	semCAP-P (ours)	semCAP-F1 (ours)	GPT-R (Nair)	GPT-P (Nair)	GPT-F1 (Nair)
CAP	**0.694**	0.212	0.256	**0.632**	**0.795**	**0.695**
CAP+Event	0.570	0.181	0.214	0.578	0.773	0.649
CAP-only	0.650	**0.269**	**0.277**	0.331	0.607	0.412
CAP+Event-only	0.141	0.077	0.065	0.132	0.288	0.167

Table 12: Transcript-free ablations (sectioned template; means). CAP+Event-only is a negative-control-style condition; CAP-only remains substantially stronger, indicating that proposition-level content can carry useful grounding signal even without transcript phrasing.

Interpretation. CAP-only keeps substantial grounded signal, but without transcript phrasing it underperforms on note-level quality. CAP+Event-only is much weaker, indicating that the current compact bundle plan is not yet sufficient as a standalone generation substrate.

F Inference Cost Analysis

To quantify the practical overhead of structure-first generation, we measure latency and token usage on a 30-case stratified subset using the same backbone model and sectioned-note template as the main experiments. We compare the deployment-time generation paths for Direct ($X \to Y$), CAP ($X \to C \to Y$), and CAP+Event ($X \to C \to E \to Y$). Reference-note CAP extraction is excluded because it is used only for offline evaluation, not deployment-time generation.

As shown in Table 13, the multi-stage pipelines increase both latency and token usage relative to Direct. CAP+Event adds only modest overhead over CAP, suggesting that most of the additional cost comes from CAP extraction and scaffold-conditioned rendering rather than event consolidation. These results quantify the current engineering cost of source-grounded structure-first generation and motivate future work on more efficient CAP induction, caching, and lightweight consolidation.

Case	Boundary regime	CAP snippet	CAP+Event snippet	Takeaway
D2N136	Organization gain (with residual risk)	"Continue metformin 500 mg twice daily" but "follow-up in two weeks" was flagged as unsupported by a reviewer.	Knee-pain problem and action items are grouped coherently (autoimmune labs + PT + meloxicam), but additional plan details (e.g., lipid profile, BP-cuff request) may over-extend beyond core focus.	Bundling can improve problem-level structure while still requiring tighter action filtering.
D2N203	Mixed trade-off	Captures vision-change workup and monthly Lucent injections, but carries restaurant narrative detail into HPI.	Maintains coherent AMD plan and follow-up cadence, yet can still retain low-value contextual detail from dialogue.	Consolidation helps organization, but noise suppression is still case dependent.
D2N010	Boundary failure	Hallucinates infection-heavy interpretation (Lyme/strep emphasis) and misses key CHF-related nuance in reviewer comments.	Over-compressed plan with omission risk (reviewers flagged missing hyperglycemia severity/right-leg edema and orthopnea-related nuance).	When evidence is noisy and multi-threaded, current event consolidation can under-specify critical findings.

Table 16: Clinician-reviewed boundary examples with generated-note snippets (sectioned template). Green marks clinically useful preservation; red marks reviewer-flagged or boundary-prone behavior.

G Boundary Analysis for Event Consolidation

Stratifying by bundle count. We stratify CAP+Event outputs by `event_plan_count` as a diagnostic proxy for consolidation strength (not a gold problem count). Table 15 shows that two-bundle plans are relatively safer, while one-bundle collapse is most associated with recall loss and under-generation.

`event_plan_count`	n	ΔsemCAP-R	ΔPDSQI	ΔPDSQI(org)	ΔF-CL(A+B+C)
1	144	−0.133	−0.093	−0.111	−0.191
2	37	−0.096	+0.039	−0.054	+0.027
3+	16	−0.110	−0.080	−0.062	−0.203

Table 15: Boundary analysis for event consolidation on the 197-case cohort (sectioned template). Values are mean deltas for CAP+Event relative to CAP. Bundle count is taken from the consolidation plan produced by cap_event.

Representative cases. Table 16 lists representative cases illustrating three regimes: (i) clear gains where consolidation improves organization and faithfulness (e.g., D2N207), (ii) trade-offs where organization improves but grounding recall drops (e.g., D2N190), and (iii) failure modes where consolidation under-generates or over-compresses, yielding large drops in semCAP-R and downstream quality (e.g., D2N153/D2N158). To make this concrete for readers, Table 16 additionally shows snippet-level highlights from the three clinician-reviewed cases (D2N136/D2N203/D2N010), marking preserved content in green and problematic rendering in red.

Case	Bundles	Regime	ΔsemCAP-R	ΔPDSQI	ΔPDSQI(org)	ΔF-CL(A+B+C)
D2N207	2	gain	+0.218	+2.286	+3.000	+1.000
D2N136	1	gain (org)	+0.176	+0.714	+1.000	−0.125
D2N190	2	trade-off	−0.178	+1.286	+1.000	+1.625
D2N205	1	failure	−0.886	−1.571	−2.000	−2.125
D2N153	1	failure	−0.714	−2.595	−4.000	−2.500
D2N158	1	failure	−0.675	−2.572	−3.000	−3.125

Table 16: Representative boundary cases for CAP+Event vs. CAP on the sectioned template; deltas are CAP+Event minus CAP.

H Extended Qualitative Example: D2N008

D2N008 is a multi-condition encounter in which the patient presents with fatigue and dizziness in the setting of low hemoglobin, background CHF, edema, salt-intake-related fluid retention, and a downstream GI workup plan. The main qualitative challenge is not merely whether these facts are mentioned, but whether they are organized around the correct primary problem. In our reading of the transcript, the clinically dominant structure is: *primary problem = anemia; evidence = fatigue, dizziness, low hemoglobin; context = CHF and edema; plan = anemia profile plus gastroenterology referral for endoscopy/colonoscopy*. This makes D2N008 a useful example of why proposition extraction and event construction are both needed.

The intermediate CAP representation also reveals a limitation that is useful to discuss explicitly. It captures the major signals well—fatigue, dizziness, swelling, low hemoglobin, prior CHF, and GI follow-up planning—but it also includes duplicated symptom propositions and weakly anchored low-value findings such as nasal congestion and seasonal allergies. This supports an important design conclusion: proposition-level extraction is necessary for faithful generation, but proposition normalization alone is not sufficient for producing a clinically prioritized note. A higher-order bundling step is particularly helpful when the encounter interleaves a primary workup problem with background chronic disease, historical detail, and peripheral conversational content.

Table 17: Dialogue snippets and core clinical challenges for D2N008.

Case	Transcript Key turns (abridged)	Clinical challenge
D2N008	U5: [patient] over the past couple of months , i've been really **tired** and **dizzy** . . . . U16: [doctor] okay , all right . so , you know , let's talk a little bit about that colonoscopy . i know you had a colonoscopy about three years ago and that showed that you had some mild **diverticuli- diverticulosis** . um , no issues since then ? U13: [patient] no , no weight loss or passing out . . . U19: [patient] um , it's been more of a struggle recently . i've been traveling a lot . i went up to vermont , um , to go , um , explore the mountains . and along the way i stopped at , you know , . . . i've noticed some **swelling** in my , my legs . . . U27: [patient] uh , i've had a little bit of **nasal congestion** just because with the seasons changing , i , i get seasonal allergies . but everything else has been okay . U32: [doctor] . . . physical examination looks fine . so , on your heart exam , i do hear a three out of six systolic ejection murmur , which we've heard in the past , okay ? and on your lower extremities , i do notice some trace to one plus **pitting edema** in your ankles , which is probably from the salt intake , okay ? U37: [doctor] . . . for you first problem this new anemia , uh , i wanna go ahead and send off some more labs and **anemia profile** , just to see exactly what type of anemia we're dealing with . i also wanna go and refer you back to the **gastroenterologist** for another evaluation , okay ? hey , dragon , order referral to gastroenterology . so , they're gonna do , uh , probably do an **endoscopy** and another **colonoscopy** on you . um , but again , i wanna send off those labs just to make sure that it's not something else , okay ?	• **Distributed evidence**: anemia workup depends on symptoms, prior GI history, edema, and follow-up plan scattered across the visit. • **Source grounding under noise**: low-salience seasonal-allergy content co-occurs with clinically central anemia evidence. • **Problem–plan linkage**: the anemia problem must remain linked to anemia profile, GI referral, endoscopy/colonoscopy, and medication adjustments. • **Avoiding over-attribution**: side details such as allergies should not dominate the assessment.

Table 18: Qualitative comparison on D2N008. This case illustrates distributed anemia-related evidence and a plan that must remain linked across symptoms, prior GI history, exam findings, and follow-up actions. CAP-based methods improve structural coherence, while Direct remains competitive as a strong baseline.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Low hemoglobin levels. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months, exacerbated by recent travel. She notes increased swelling in her ankles and reports difficulty keeping up with her usual activity level. She had a colonoscopy three years ago that revealed mild diverticulosis with no complications. She previously experienced kidney stones two years ago, with no recent flares. She denies any recent blood in her stools or melena. **Findings** Physical examination reveals trace to one plus pitting edema in the ankles. The patient's blood pressure is within normal limits, and her heart rate is appropriate. A three out of six systolic ejection murmur is heard on cardiac auscultation. The echocardiogram demonstrates a left ventricular ejection fraction of 45% and moderate mitral regurgitation. Hemoglobin is 8.2 g/dL. **Assessment** The patient presents with anemia, likely secondary to chronic disease or other underlying causes. Her congestive heart failure and recent diuretic use may contribute to her symptoms. **Plan** I am ordering an anemia profile to further evaluate the cause of the anemia. I am also ordering a referral to gastroenterology for an endoscopy and repeat colonoscopy. I will initiate lasix 40 milligrams once daily, continue toprol 50 milligrams daily, and lisinopril 10 milligrams daily. I will instruct the patient to monitor her weight daily and call if there is any weight gain. I will also recommend a scale to assist with monitoring sodium intake.	A strong direct baseline: it covers much of the visit correctly. Its weakness is that the overall rationale is less explicitly scaffolded, making it harder to see how distributed evidence was organized.
Cluster2Sent	**Chief Complaint** The patient reports fatigue and dizziness. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months. She denies any recent trauma. She reports some mild nasal congestion related to seasonal allergies. She denies fever, chills, or sore throat. She denies any weight loss or loss of appetite. She reports some swelling in her lower extremities, primarily in her ankles. She has no issues since her previous colonoscopy three years ago. **Findings** The patient is a 49-year-old female with a past medical history of congestive heart failure, kidney stones, and a prior colonoscopy. Her vital signs are stable, with blood pressure 132/78 and heart rate 78. A three out of six systolic ejection murmur is auscultated. The echocardiogram reveals a left ventricular ejection fraction of 45%, moderate mitral regurgitation, and trace to one plus pitting edema in the ankles. Hemoglobin is low. **Assessment** The patient presents with anemia, likely secondary to chronic disease and requiring further investigation. She also continues to have symptoms consistent with congestive heart failure. **Plan** I am going to order additional labs, including a complete blood count and iron studies. I will also refer her back to the gastroenterologist for an endoscopy and repeat colonoscopy. I will initiate lasix 40 milligrams once daily and continue toprol 50 milligrams daily. I will also consider a diuretic increase if symptoms persist.	Reasonably faithful, but flatter and less explicit about which evidence belongs to the anemia workup versus background chronic conditions.
MEDSUM-ENT	**Chief Complaint** *[empty]* **History of Present Illness** *[empty]* **Findings** *[empty]* **Assessment** *[empty]* **Plan** *[empty]*	Under-generated sectioned output (empty skeleton), limiting its usefulness as a rendering scaffold for long multi-condition narratives in our prompt-based reimplementation.

Method	Generated note (sectioned template, full)	Comment
CAP	**Chief Complaint** Low hemoglobin level. **History of Present Illness** The patient reports fatigue and dizziness over the past couple of months, particularly with increased exertion. She notes swelling in her ankles and reports difficulty with salt intake, particularly while traveling. She had a colonoscopy three years ago that revealed mild diverticulosis with no issues since. She also has a history of kidney stones two years ago with no recent flares. She reports seasonal allergies and nasal congestion. She denies any weight loss or loss of appetite, or episodes of passing blood in the stool or urine. **Findings** The patient's blood pressure is within normal limits. Heart rate is appropriate. Physical exam reveals trace to one plus pitting edema in the ankles. A systolic ejection murmur is heard. The doctor notes nasal congestion. **Assessment** The patient's low hemoglobin level is concerning and requires further investigation. Her congestive heart failure, diverticulosis, and seasonal allergies are contributing factors. **Plan** The doctor plans to order an anemia profile and a referral to the gastroenterologist for an endoscopy and colonoscopy. The doctor also orders lasix 40 milligrams once daily, toprol 50 milligrams daily, and lisinopril 10 milligrams a day. The doctor recommends monitoring weight daily and calling if it increases. The gastroenterologist will perform an endoscopy and colonoscopy.	CAP preserves the main anemia-centered plan well, but still over-carries peripheral details such as seasonal allergies into the note.
CAP + Event	**Chief Complaint** Low hemoglobin levels and fatigue. **History of Present Illness** Stephanie is a 49-year-old female with a history of congestive heart failure, kidney stones, and a prior colonoscopy. She presents today with complaints of fatigue and dizziness over the past couple of months, noting that these symptoms have worsened recently, particularly with travel. She reports increased swelling in her ankles and has been struggling to maintain a low-salt diet. She had a colonoscopy three years ago that revealed mild diverticulosis with no further issues. She also had kidney stones a couple of years ago, with no recent flares. She reports occasional nasal congestion related to seasonal allergies but no other significant symptoms, including fever, chills, sore throat, or cough. She denies any weight loss or loss of appetite and has not experienced any episodes of fainting or dark stools. **Findings** Physical exam reveals blood pressure 128/78 mmHg, heart rate 72 bpm, and trace 1+ pitting edema in the ankles. A three out of six systolic ejection murmur is heard. The echocardiogram shows a left ventricular ejection fraction of 45% and moderate mitral regurgitation. Hemoglobin is 8.2 g/dL. **Assessment** The patient presents with anemia, likely secondary to chronic disease and lifestyle factors. Her congestive heart failure and recent travel contribute to her symptoms. **Plan** I am ordering an anemia profile to determine the etiology of the anemia. I will also refer her back to the gastroenterologist for an endoscopy and repeat colonoscopy. I will initiate lasix 40 mg daily, continue toprol 50 mg daily, and lisinopril 10 mg daily. I will also order a referral to gastroenterology for further evaluation. I will instruct the patient to monitor her weight daily and call if there is any significant increase.	CAP+Event most clearly organizes the visit around the anemia problem and its downstream workup, though it can still introduce mild over-explanation in narrative form.

I Extended Qualitative Example: D2N026

D2N026 is our main safety-critical supporting case. The encounter is structurally simple compared with D2N008, but it contains a clinically important contradiction that occurs late in the dialogue: the patient first denies blood thinners and later states that he is taking Eliquis. Because the downstream plan includes cervical MRI and a tentative epidural, recovering the *final grounded medication state* is more important here than broad narrative fluency. This case therefore illustrates a second recurring failure mode in clinical dialogue summarization: brittle final-state tracking under long-range dependency. Table 19 summarizes the key dialogue turns and why this failure mode is safety-critical for downstream planning.

Table 19: Dialogue snippets and core clinical challenges for D2N026.

Case	Dialogue evidence (abridged)	Clinical challenge
D2N026	**Key turns** U9: "the thing started about **two weeks ago**" U19–U27: arm/hand pain extends down the arm; **hand weakness** and grip difficulty U49–U55: cervical source, **MRI** needed, tentative **epidural** plan U57: "do you take any **blood thinners**?" U58: "no, I don't" U59: "I am taking **Eliquis**, though" U65: "my hand is still bothering me" U66–U68: **MRI of neck + C7–T1 epidural** plan maintained	• **State update conflict**: earlier negation → later correction. • **Medication safety**: anticoagulant status matters. • **Primary problem grounding**: cervical / arm–hand pain. • **Plan preservation**: neck MRI + epidural plan.

Table 20: Qualitative comparison on D2N026. Direct generation collapses to an unrelated problem, while Cluster2Sent keeps the wrong earlier medication state. CAP-based methods better preserve the later clinically actionable Eliquis update together with MRI and epidural planning.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Right shoulder pain **History of Present Illness** The patient presents with right shoulder pain that began approximately 3 days ago after lifting a heavy box. He describes the pain as sharp and constant, exacerbated by overhead movements. He denies any recent trauma. He has tried over-the-counter ibuprofen with minimal relief. He reports difficulty reaching behind his back. He has no prior history of shoulder problems. He is currently taking lisinopril and metformin. **Findings** Physical exam reveals tenderness to palpation over the right rotator cuff. Range of motion is limited by pain. Neurological exam is intact. **Assessment** Rotator cuff strain is suspected. **Plan** Continue with ice and elevation. Initiate physical therapy. Consider a rotator cuff ultrasound to rule out a tear.	Catastrophic topic drift. The note misses the cervical MRI, epidural planning, gabapentin continuation, and anticoagulant update altogether.
Cluster2Sent	**Chief Complaint** Left arm pain **History of Present Illness** The patient reports left arm pain, initially attributed to a possible pinched nerve by a previous provider. He has been taking gabapentin and a neck pill, and reports using numbing medicine. He denies taking blood thinners. He reports some mild strain. **Findings** The patient has mild cervical restriction on examination. Cervical x-ray reveals significant disc degeneration at C5–C6 and to a lesser extent C4–C5 and C3–C4. MRI of the neck is planned to assess for nerve compression. Physical exam reveals no apparent distress. **Assessment** Suspect cervical radiculopathy with possible contributing peripheral neuropathy. The patient's presentation is consistent with cervical disc degeneration and radicular symptoms. **Plan** Order a neck MRI to assess for nerve compression. Tentatively plan for a left C7–T1 epidural, pending MRI results. Continue the patient on gabapentin and advise him to continue his home exercise program. Schedule a follow-up appointment in four weeks if symptoms persist.	Better problem anchoring than Direct, but it preserves the *wrong final state* by carrying forward the earlier blood-thinner denial instead of the later Eliquis correction.
MEDSUM-ENT	**Chief Complaint** *[empty]* **History of Present Illness** *[empty]* **Findings** *[empty]* **Assessment** *[empty]* **Plan** *[empty]*	In our prompt-based reimplementation, the sectioned renderer under-generated for this case, producing an empty skeleton. (The SOAP template produces a non-empty output but still struggles to reconcile the late Eliquis update.)
CAP	**Chief Complaint** Left arm pain. **History of Present Illness** The patient reports hand pain, which began approximately two weeks ago. He describes the pain as worsening at night and initially experienced it when waking at 2:00 AM. He reports that applying ice provided temporary relief. The patient also notes difficulty gripping things and reports weakness in his hand. He previously injured his hand and has a history of leg issues and knee issues, and is one percent disabled. He reports taking eliquis and gabapentin. He previously received numbing medicine and oral steroids. **Findings** Physical exam reveals mild cervical restriction with limited extension, and tenderness to palpation in the neck. Strength is one plus bilaterally in the biceps, triceps brachii, and brachioradialis. Reflexes are normal. Examination does not reveal any specific findings in the upper extremities other than the reported hand weakness. **Assessment** The patient's symptoms are most likely due to cervical radiculopathy, possibly secondary to disc degeneration. Peripheral neuropathy is also a consideration. **Plan** I am going to order a cervical MRI to further evaluate the nerve roots and determine the extent of compression. I will also tentatively schedule a C7–T1 epidural. The patient will continue taking gabapentin and eliquis. I will schedule a follow-up appointment in four weeks to assess his progress.	CAP recovers the clinically critical later correction (**Eliquis**) and preserves the cervical radiculopathy workup and treatment plan.

Method	Generated note (sectioned template, full)	Comment
CAP + Event	**Chief Complaint** Left arm pain. **History of Present Illness** The patient reports left arm pain that began approximately two weeks ago, worsening at night. The pain initially arose suddenly and was accompanied by a feeling of discomfort, but it has since become more persistent. He describes the pain as a dull ache, and it seems to be exacerbated by certain movements. He mentions that he experienced similar symptoms in the past, which were attributed to a pinched nerve by a previous physician. He has been taking gabapentin and eliquis as prescribed. He denies any recent trauma or injury to the arm. **Findings** Physical exam reveals mild tenderness to palpation in the cervical region. Range of motion is limited in cervical flexion and extension. Strength testing reveals a one-plus grade for biceps and triceps strength bilaterally. Reflexes are normal. Cervical x-ray demonstrates significant disc degeneration at C5–C6 and C4–C5. **Assessment** The patient's symptoms are most likely due to cervical radiculopathy, possibly exacerbated by disc degeneration. Peripheral neuropathy is also a consideration. **Plan** I am ordering a cervical MRI to further evaluate the nerve roots and determine the extent of compression. We will also schedule a left C7–T1 epidural to provide targeted anti-inflammatory relief. The patient will continue with gabapentin as prescribed and follow up in four weeks to assess symptom improvement.	CAP+Event preserves the main problem–evidence–plan chain and avoids the unsafe blood-thinner denial, though some detail is compressed during rendering.

J Extended Qualitative Example: D2N024

This case is useful precisely because it is not a clean victory. The clinically dominant structure is mild scar discomfort with conservative management and deferred revision, while the lump belongs to historical context and the review-of-systems negatives should remain suppressed. As shown in Tables 21 and 22, Cluster2Sent is brittle under negation-heavy follow-up dialogue and can promote explicitly negated symptoms into positives, while MEDSUM-ENT and Direct remain strong on this visit style. CAP contributes by making temporal resolution and the conservative scar-centered plan explicit, but it can still over-carry speculative counseling content. We therefore view D2N024 as evidence that structured representations improve robustness under noisy dialogue, but do not fully solve the filtering problem when low-salience follow-up content and speculative counseling are tightly interleaved.

Table 21: Dialogue snippets and core clinical challenges for D2N024.

Case	Dialogue evidence (abridged)	Clinical challenge
D2N024	**Key turns** U8: "**no**" to **chills**, **fever**, **nausea**, **vomiting** U18: back pain **resolved** U21: **lump under left breast** was noted in March … U23: **scars were bothering her** … U33–U47: conservative plan: **scar gel**, **massage**, photos, **revisit in 6 months**	• **Temporal resolution**: prior lump → no longer present. • **Negation-heavy follow-up**: ROS items can be falsely promoted. • **Low-salience follow-up noise**: social talk can leak into note text. • **Current problem vs. future option**: revision is future/hypothetical.

Table 22: Qualitative comparison on D2N024. This follow-up visit mixes a scar complaint with a resolved historical lump and negation-heavy review-of-systems content.

Method	Generated note (sectioned template, full)	Comment
Direct	**Chief Complaint** Breast scar discomfort following bilateral reduction mammoplasty. **History of Present Illness** The patient reports persistent discomfort with the scars from her bilateral reduction mammoplasty performed in October 2020. She initially reported irritation and sought relief with scar gel, applying it twice daily. She states the scars were bothersome and used the gel to mitigate this. She has not had any further issues with the scars since her last visit with Dr. Sanchez in March. She denies any new chills, fever, nausea, or vomiting. She reports enjoying summertime activities and her family is doing well. She has not experienced any new back pain. **Findings** Physical examination reveals well-healed incisions with no signs of infection or redness in either breast. The incisions appear widened, particularly on the left side. The patient reports subjective discomfort with the left scar. **Assessment** Postoperative scarring following bilateral reduction mammoplasty. **Plan** Schedule a follow-up appointment in six months to reassess the scars. Continue the use of mederma scar gel twice daily on the affected areas, massaging the scars. Obtain photographs of the scars for documentation and future comparison.	Strong baseline: keeps the scar follow-up as primary, preserves ROS negations, and treats the prior lump as historical/resolved.
MEDSUM-ENT	**Chief Complaint** Scar discomfort following bilateral reduction mammoplasty. **History of Present Illness** The patient reports ongoing discomfort with the scars from her previous bilateral reduction mammoplasty, specifically noting that one scar area is bothering her. She previously reported a lump in that area to her primary care physician, Ruth Sanchez, in March, but no longer feels it. She has been using scar gel twice daily to manage the scar tissue and has observed some widening of the incisions. She denies any signs of infection or redness. The patient is currently enjoying summertime and has not experienced any new symptoms or concerns. **Findings** Examination reveals well-healed incisions with no signs of infection or redness on either breast. The patient reports the scar tissue is widening. **Assessment** The patient's primary concern is persistent scar discomfort following breast surgery. There is no evidence of acute infection or complications. **Plan** The patient should continue using the mederma scar gel twice daily on the affected scar area. We will schedule a follow-up appointment in six months to reassess the scar tissue and determine if further intervention is needed. Photographs will be taken today to document the current state of the scars for future reference.	Captures temporal resolution (lump no longer present) and negation-heavy content cleanly; aligns with MEDSUM-ENT's status-centric design.
Cluster2Sent	**Chief Complaint** Patient reports concern regarding breast scars following bilateral reduction mammoplasty. **History of Present Illness** The patient reports experiencing chills, fever, nausea, and vomiting. She describes a lump located under her left breast, which she has not felt recently. She notes that the scars are bothering her and have widened slightly, particularly in a high-tension area. She has been using scar gel and massage to manage the scars. **Findings** The incisions are well-healed with no signs of infection or redness on either breast. **Assessment** The patient's reported breast lump and scar concerns require further evaluation. **Plan** Schedule a follow-up appointment in six months to reassess the scars and address any ongoing concerns. Recommend continued scar gel application and massage.	Promotes explicitly negated ROS items to positives (unsafe) and blurs the historical lump vs. current scar follow-up distinction.
CAP	**Chief Complaint** Patient reports scar discomfort and a desire for scar revision. **History of Present Illness** The patient reports a breast lump discovered during self-examination in March, which was evaluated by Dr. Sanchez. She reports that the scar from her previous reduction mammoplasty has been bothering her, and she has been using scar gel twice daily. She denies any recent trauma or fever. She reports no nausea or vomiting. She describes the scar as being slightly bothersome, and she's been massaging the scar gel. She's been using the gel for about a month. She's been using the gel twice a day. She discontinued using the gel because it was bothering her. She's been enjoying summertime and hasn't had any new problems. **Findings** Examination reveals well-healed incisions with no signs of infection or redness. The incisions appear widened. The patient denies chills, fever, nausea, or vomiting. **Assessment** The patient has scars from previous surgery and reports discomfort. There is no acute infection or signs of complications. **Plan** The doctor recommends the patient continue using scar gel twice a day and massaging the scar gel. The doctor recommends revisiting the scar in six months. The doctor will take pictures of the scars today and will revisit the scars in six months. The doctor recommends surgical excision if necessary.	Preserves key facts, but over-carries speculative counseling (revision/excision) and includes repetition/noise in HPI.

Method	Generated note (sectioned template, full)	Comment
CAP + Event	**Chief Complaint** Patient is Pamela Cook, a 36-year-old female post-bilateral reduction mammoplasty on 10-10-2020, presenting for follow-up regarding scar discomfort and a recent lump. **History of Present Illness** The patient reports being Pamela Cook, a 36-year-old female post-bilateral reduction mammoplasty on 10-10-2020. She states she is doing well and enjoys summertime. She reports that the scars here were kind of bothering me and I got scar gel. I was using it everyday, but I do n't think I need it now. She also mentioned a lump right here under her left breast that she felt during her mammogram in March, which was noted by Dr. Ruth Sanchez. However, she hasn't felt it since then. She discontinued using scar gel. She denies any new chills, fever, nausea, or vomiting. **Findings** Physical examination reveals well-healed incisions with no signs of infection or redness on either breast. The patient's back pain is currently not a concern. **Assessment** The patient's presentation is consistent with post-operative scar discomfort and a previously identified, now absent, palpable lesion, warranting continued monitoring. **Plan** The doctor recommends revisiting the scar in six months. The patient can keep using mederma scar gel twice a day. The patient can put scar gel on other scars. The patient should massage the scar gel. The doctor will take pictures of the scars today. The doctor will revisit the scars in six months.	Better separates historical vs. current status and keeps the conservative follow-up plan; still contains some templated repetition (name/date).

K Clinician Deep Review Protocol and Summary

We conducted a focused deep review on three sectioned-note cases: D2N136 (organization gain), D2N203 (mixed trade-off), and D2N010 (boundary failure). Two clinicians independently rated Direct, Cluster2Sent, CAP, and CAP+Event using a checklist co-designed with the reviewers to directly probe the failure modes in our Motivation and Research Questions (Figure 5). The checklist contains eight 1–5 Likert items spanning (A) content preservation, (B) clinical state fidelity, and (C) problem-oriented organization, plus safety flags (omission/hallucination/major error; yes/no). We also collected a *clinical usability* rating as an auxiliary item and report it separately (not included in the A+B+C aggregate). Overall, the three cases cover (i) an organization gain, (ii) a preservation–organization trade-off, and (iii) a boundary failure where all systems show clinically meaningful weaknesses.

Reviewer summaries.

- **Clinician A:** On close reading, notes can still contain omissions/hallucinations and occasional section-miscategorization even when they look broadly plausible.
- **Clinician B:** No severe errors were observed; differences were largely preference-level, and conclusions are limited with two reviewers.

Per-case clinician sheets. Tables 24–25 summarize the raw scoring sheets from our two clinicians (A and B) with brief English translations of their key comments. Narrative notes made Problem–Evidence Linkage difficult to rate in some cases (N/A), consistent with our decision to treat this item cautiously in the main quantitative analysis.

Clinician deep-review checklist (1–5 Likert; 1=poor, 5=excellent).

A. Content Preservation

1 (C) **Core clinical information preserved:** Are clinically salient facts from the dialogue retained without major omission?

2 (N) **Noise suppression:** Are small talk, repeated questions, and peripheral context avoided in the note?

3 (Ly) **Lay-expression fidelity:** Is patient language neither copied verbatim nor over-abstracted, preserving symptom nuance (severity/uncertainty)?

B. Clinical State Fidelity

4 (F) **State update / final-state fidelity:** Are late corrections and resolved uncertainty reflected in the final state (unknown→resolved; negation→positive; current med vs. intended change)?

5 (T) **Temporality preservation:** Are past history, current status, and future plan kept distinct?

C. Problem-oriented Organization

6 (Lk) **Problem–evidence linkage:** Are symptoms/exam/tests/history linked under the correct clinical problem?

7 (SP) **State–plan separation:** Are current state/current meds separated from new orders/prescriptions/follow-up plans?

8 (SO) **Section / note organization:** Are items placed in appropriate SOAP/sectioned sections?

Safety flags (Y/N): O=clinically meaningful omission; H=clinically concerning hallucination; M=major error.

Auxiliary: U=clinical usability (1–5; reported separately; excluded from the A+B+C faithfulness aggregate).

Figure 5: Clinician deep review rubric co-designed with the clinician reviewers to directly probe the failure modes in Motivation and Research Questions. Abbreviations match Tables 24–25.

Method	A Content	B State	C Org	Mean (A+B+C)	Clinical Usability
Direct	3.78	**5.00**	**4.89**	4.54	3.33
Cluster2Sent	4.11	**5.00**	4.72	4.63	3.67
CAP	**4.19**	**5.00**	4.72	4.62	3.33
CAP+Event	4.17	**5.00**	4.83	**4.77**	**4.00**

Table 23: Clinician deep review summary (mean across three ACI-Bench cases × two clinicians). Scores are 1–5 Likert ratings aggregated by rubric category (Figure 5); N/A items are excluded. Direct is highest on overall organization (C) in this small sample; among the two clustering/bundling approaches, CAP+Event outscored Cluster2Sent on organization and the A+B+C mean.

Case	Method	C	N	Ly	F	T	Lk	SP	SO	U	O/H/M	Key note (translated)
D2N136	Direct	3	5	N/A	N/A	5	N/A	5	5	3	Y/Y/–	Omits negated ROS items (e.g., vomiting/chest pain/dyspnea) and hallucinates a 4-week follow-up; state-update and problem–evidence linkage were not applicable to rate in this case.
D2N136	Cluster2Sent	3	5	N/A	N/A	5	N/A	5	4	3	Y/–/–	Best plan capture among the compared methods for this case, but omissions were still flagged.
D2N136	CAP	4	5	N/A	N/A	5	N/A	5	5	4	–/Y/–	Hallucinates a 2-week follow-up appointment.
D2N136	CAP+Event	3	3	N/A	N/A	5	N/A	5	3	3	Y/Y/–	Omits continued metformin (500mg BID) from the plan and hallucinates a 4-week follow-up; misses prior rotator cuff repair and over-includes low-value narrative; BP is missectioned.
D2N203	Direct	4	2	4	N/A	5	N/A	5	5	3	–/–/–	Includes conversational filler (e.g., "I will return in a few minutes") that should not appear in a clinical note.
D2N203	Cluster2Sent	5	4	3	N/A	5	N/A	5	4	4	–/Y/–	Hallucinates an "eye needle-stick" history not supported by the dialogue.
D2N203	CAP	5	3	4	N/A	5	N/A	5	5	4	–/–/–	No major issues were noted in the clinician sheet for this case/method.
D2N203	CAP+Event	5	3	4	N/A	5	N/A	5	5	4	–/–/–	No major issues were noted in the clinician sheet for this case/method.
D2N010	Direct	3	5	5	5	5	N/A	5	5	3	–/Y/–	Hallucinates a Lyme titer and a 4-week follow-up; overfocuses on side-sleeping and misses the key orthopnea-negative point (lying flat is not problematic).
D2N010	Cluster2Sent	3	4	4	5	5	N/A	5	5	3	Y/Y/–	Omits severity of hyperglycemia (blood glucose in the 300s); hallucinates/incorrectly summarizes rapid strep treated with amoxicillin and misstates alpha-gal "certainty."
D2N010	CAP	4	5	4	5	5	N/A	5	5	3	Y/–/Y	Omits right-leg edema in a CHF context; major interpretation error around sleep/orthopnea (misses that lying flat is fine).
D2N010	CAP+Event	3	5	4	5	5	N/A	5	5	3	Y/Y/–	Omits hyperglycemia severity and right-leg edema; incorrectly implies the patient avoids lying flat (dialogue indicated no orthopnea).

Table 24: Clinician A deep review sheet (translated). Scores are 1–5 Likert ratings: C=Core information, N=Noise suppression, Ly=Lay-expression fidelity, F=Final-state fidelity, T=Temporality preservation, Lk=Problem–Evidence linkage, SP=State–Plan separation, SO=Section organization, U=Overall clinical usability. O/H/M indicates whether the reviewer flagged omission/hallucination/major error (Y/–).

Case	Method	C	N	Ly	F	T	Lk	SP	SO	U	O/H/M	Key note (translated)
D2N136	Direct	2	5	3	5	5	3	5	5	3	–/–/–	Incorrect patient sex; includes an irrelevant statement about lower-extremity strength.
D2N136	Cluster2Sent	2	5	4	5	5	3	5	5	3	–/–/–	Captures the chief complaint, patient sex, and diabetes history.
D2N136	CAP	2	5	5	5	5	5	3	2	1	–/–/–	Includes content in Assessment that does not belong there.
D2N136	CAP+Event	5	5	5	5	5	5	5	5	5	–/–/–	Best overall among the four methods for this case.
D2N203	Direct	2	2	5	5	5	5	5	5	3	–/–/–	Italian-restaurant detail is unnecessary; HPI is overly long and could be condensed.
D2N203	Cluster2Sent	4	5	5	5	5	5	5	5	4	–/–/–	No additional comment provided in the sheet.
D2N203	CAP	4	2	5	5	5	5	5	5	3	–/–/–	Carries over the irrelevant "Italian restaurant" episode.
D2N203	CAP+Event	3	4	5	5	5	5	5	5	4	–/–/–	Overall clean, but seems to omit many findings.
D2N010	Direct	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	Cluster2Sent	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	CAP	4	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.
D2N010	CAP+Event	5	5	5	5	5	5	5	5	5	–/–/–	No additional comment provided in the sheet.

Table 25: Clinician B deep review sheet (translated). Column definitions match Table 24. (This sheet did not include omission/hallucination/major-error flags for most rows, so O/H/M is shown as "–" throughout.)

Segmentation Matters: Exploring LLM-Based Strategies for Temporal Clinical Event Identification in Oncology Reports

Cristiano Bellucci[4], Francesco Madeddu[2,3], Chiara Iacomini[1], Carlotta Masciocchi[1], Stefano Patarnello[1], Massimo Bernaschi[2], Mario Santoro [2], Livia Lilli [1]

[1] Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

[2] Istituto per le Applicazioni del Calcolo "Mauro Picone", Italian National Research Council

[3] Department of Computer, Control and Management Engineering, Sapienza University of Rome

[4] KeyBiz srl

livia.lilli@policlinicogemelli.it

Abstract

Processing unstructured clinical narratives remains a major challenge in medical Natural Language Processing (NLP), particularly when critical information is embedded within lengthy and heterogeneous reports. Clinical notes often describe key diagnostic and therapeutic events through a verbose narrative, making automatic event identification difficult. In this work, we frame the identification of clinical events as a text segmentation task. We conduct a comparative study of three segmentation strategies applied to oncology reports: (i) a fully regex-based approach, (ii) a cascaded regex–LLM pipeline, and (iii) the same cascade architecture augmented with a recovery mechanism to mitigate LLM rephrasing. Segmentation quality is evaluated using complementary structural metrics (Pk, WindowDiff, Boundary Similarity, Segment Count Accuracy, and Text Overlap IoU), and its impact is also observed on downstream segment tagging, performed to identify the corresponding event type (e.g. surgery, biopsy, imaging, treatment, laboratory). The results demonstrate the high potential of LLM-based approaches, particularly in preserving semantic coherence within segments and generalization on new data sources. However, regex-based segmentation achieves higher performance according to structural segmentation metrics, also leading to better downstream clinical event identification. In general, these results highlight the critical role of context-adaptive high-quality segmentation strategies in the structuring of verbose clinical narratives and in the accurate identification of key patient events.

1 Introduction

Processing unstructured clinical narratives remains a major challenge in medical Natural Language Processing (NLP). Clinical reports are typically lengthy and unstructured documents that aggregate heterogeneous information in a verbose narrative, including key events like diagnostic findings, therapeutic interventions, laboratory results, and follow-up plans. This structural complexity makes them difficult to handle within downstream NLP pipelines, where models often require well-defined and coherent input units.

To address this issue, preprocessing strategies are commonly employed to make clinical texts more workable. In particular, text segmentation can support downstream tasks by dividing reports into smaller and semantically coherent units, facilitating more targeted processing. However, chunking techniques used in transformer-based and LLM architectures are predominantly length-driven rather than meaning-driven (Jaiswal and Milios, 2023). As a consequence, they may split text at arbitrary boundaries, disrupting semantic coherence and potentially separating information that belongs to the same clinical event, or they may fail to identify semantically distinct sub-sections, relative to two different events. In this direction, Lilli et al. (2026) showed that content-aware segmentation of clinical reports can improve downstream information extraction by enabling content-specific prompting strategies for different clinical event types. In that work, the primary focus was not the segmentation task itself, but the evaluation of extraction performance under different prompt configurations tailored to the segmented report content. Within this framework, segmentation represented a key preprocessing step for the extraction architecture and was implemented through a preliminary regex-based approach.

In this study, we reframe text segmentation as a semantics-driven task aimed at identifying clinically meaningful event units, and not only reducing input length. Focusing on the clinical-history section of oncological reports, we seek to automatically detect and isolate key temporal events (such as imaging examinations, surgeries, laboratory assessments, biopsies, and treatment lines) that define the patient's clinical trajectory. These

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 595–604
July 3-4, 2026 ©2026 Association for Computational Linguistics

events are typically embedded within chronological narrative text, where explicit structural markers may be inconsistent or absent, making boundary detection particularly challenging. To address this problem, we conduct a comparative evaluation of three segmentation strategies: (i) a fully rule-based approach relying on regular expressions and structural cues, (ii) a regex–LLM pipeline that refines rule-based segments through semantic modeling, and (iii) the hybrid regex-LLM framework incorporating a recovery mechanism to mitigate LLM-induced boundary errors and hallucinations. In addition to intrinsic segmentation metrics, we evaluate each strategy based on its downstream event tagging performance, demonstrating that segmentation quality directly affects the accuracy of event type classification. Through this analysis, we highlight the complementary strengths of deterministic rules and LLM-based approaches, and demonstrate the importance of high-quality text segmentation for accurately identifying key events within a patient's clinical trajectory. The study framework is shown in Figure 1.

2 Background

Early approaches to document structuring relied heavily on rule-based systems to identify section boundaries. Pioneering works like SecTag (Denny et al., 2009) utilized terminological headers and line formatting to detect document sections, a methodology that remains influential in modern pipelines such as medspaCy (Eyre et al., 2021). Although these approaches effectively handle document structures, they often lack the granularity required to tell apart distinct clinical events within narrative-heavy sections like the clinical history, where explicit headers are frequently replaced by chronological descriptions and date references.

The granular identification of such events is closely linked to temporal information extraction. Major initiatives like the i2b2 Shared Task (Sun et al., 2013) and the THYME corpus proposed for use in a SemEval 2015 task (Styler IV et al., 2014) established the standard for extracting events, temporal expressions, and their relationships. These tasks were approached mostly with ML supervised methods, which typically require large, annotated corpora (scarce in real-world settings).

Furthermore, recent work by Lilli et al. (2025b) shows how sentence boundary detection helps text segmentation using both traditional and LLM-based methods, including SaT (Frohmann et al., 2024), PySBD (Sadvilkar and Neumann, 2020), Stanza (Qi et al., 2020) and NLTK (Bird, 2006). The benchmark analysis shows that even sentence-level boundary detection can significantly enhance downstream clinical report classification and critical event identification.

Building on the above findings, text segmentation emerges as a crucial and actively explored task in clinical NLP. However, despite its importance, it remains challenging due to the variability and complexity of clinical narratives. With recent advances in NLP, many clinical text processing tasks are increasingly being addressed using LLM-based approaches, which have enabled substantial progress in areas such as clinical information extraction and medical text generation (Yang et al., 2022; Agrawal et al., 2022; Lilli et al., 2025a). Recent comparative studies, such as CNSight (Surana et al., 2025), have evaluated rule-based, transformer, and LLM-based approaches for clinical note segmentation, similarly finding that rule-based methods can be robust. While with CLINES (Yang et al., 2025b), the complex space of clinical text segmentation for semantical analysis is further explored.

Despite their strong semantic understanding, LLMs remain prone to hallucinations, often producing plausible yet clinically inaccurate information (Ji et al., 2023). To mitigate this limitation, hybrid approaches have been proposed that combine the deterministic precision of regex with the reasoning capabilities of LLMs (Kleinlein et al., 2025). Crucially, a major challenge in LLM-based segmentation is ensuring that the generated output aligns exactly with the legally valid source text. Recent efforts such as LangExtract (Goel, 2026) and SafePassage (Barrow et al., 2025) focus heavily on grounding LLM outputs back to source spans. The closest work to our proposed recovery approach is MedSlice (Davis et al., 2025), which addresses LLM boundary misalignment in section-level segmentation using Levenshtein-based fuzzy matching. Building upon this line of research, our work introduces an extended recovery mechanism that leverages full-segment fuzzy matching rather than anchor n-grams, incorporates gap-filling between aligned segments, and implements a global rollback strategy when alignment fails.

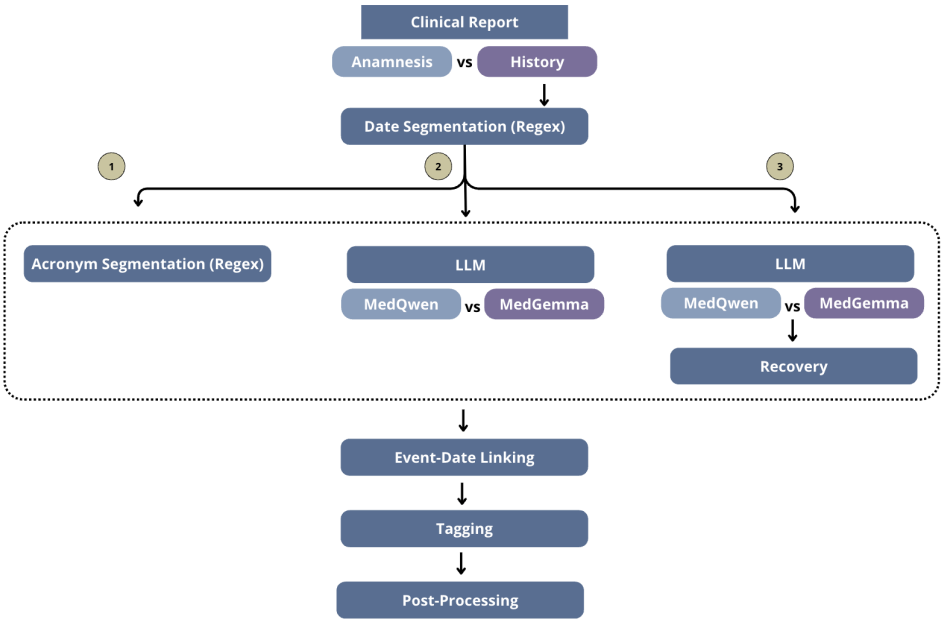


Figure 1: Study workflow with the 3 segmentation strategies: (1) regex-only, (2) regex + LLM, (3) regex + LLM + recovery.

3 Method

3.1 Dataset

The dataset consists of anonymized, real-world Italian ovarian oncology reports from multidisciplinary tumor board sessions at the Gemelli Hospital in Rome, where clinicians discuss patient cases and decide on follow-up and care. The documents are typically extensive and semi-structured. Each report opens with an anamnestic summary of the patient's current status, followed by a clinical history section. This history is organized as a sequence of events (such as surgeries, diagnostic imaging, or systemic treatments) which are generally prefaced by an explicit date or a functional header that provides a brief summary of the event type (e.g., *TC total body* for an imaging event). The primary objective of this study is to leverage this structure to automatically isolate these clinical events, transforming the narrative history into independent textual segments.

3.2 Text Segmentation

The study considers three text-segmentation strategies applied exclusively to the clinical history section of the reports: (1) a regex-only method, (2) a cascaded regex–LLM pipeline, and (3) the same cascaded pipeline augmented with a recovery step.

3.2.1 Regex-only

The first approach identifies event boundaries by detecting date expressions strictly at the beginning of a sentence or line-initial position. Supported date formats include `dd/mm/yyyy`, `dd/mm/yy`, month–year expressions (e.g., `month, year`), and `dd/mm` without year. Dates preceded solely by plain text are ignored to avoid spurious intra-sentential matches.

After date-based segmentation, further refinement is performed by analyzing patterns that may correspond to event titles or section headers. This step is designed to capture clinical events that lack an explicit date but are introduced by a functional header. To this purpose, a domain-specific dictionary of acronyms and lexical variants used to denote clinical events is employed to support boundary detection. Table 1 reports the Italian patterns used to identify event types. We focus on the key clinical events in the oncological patient trajectory, namely: imaging (including instrumental examinations such as CT, MRI, X-ray, and ultrasound), surgery, laboratory tests (where biomarker values are typically reported), treatment, and biopsy. Finally, flow 1 in Figure 1 represents the Regex-only segmentation.

3.2.2 Regex + LLM

In the two-stage cascade approach, the clinical-history section is first segmented into event units

using the date-based regex rules described above. Each resulting segment is then provided as input to an LLM, to which it is explicitly requested to identify further sub-segments corresponding to key clinical events, based on a predefined list of domain-specific acronyms. These acronyms are used as anchors to detect meaningful intra-segment boundaries that typically introduce events not detected through the above date-regex approach. The LLM returns a structured JSON output, which is automatically parsed to extract the corresponding facts and store them in a normalized representation for a granular segmentation. Flow 2 in Figure 1 represents this strategy. The prompt used is the following (we report the English version for better understanding):

```
Segment the following clinical text into separated
facts based ONLY on medical acronyms.

ACRONYMS that indicate a new segment:
- Imaging: {list of acronyms}
- Surgery: {list of acronyms}
- Laboratory: {list of acronyms}
- Therapy: {list of acronyms}
- Biopsy: {list of acronyms}
- Other: {list of acronyms}

Provide the output ONLY in JSON format without
preambles or explanations.
The facts must correspond exactly to the segments
     of
the clinical report.
Example output: {fact1: <text of fact 1>, ..}.

If there are no acronyms, return a JSON with a
single fact: {fact1: <text of fact 1>}.

TEXT: {input text}
```

Event type	Acronyms / keywords
Imaging	TC; TAC; TCTAP; TCTB; RM; RMN; PET; ECO; Ecografia; ETG; ECOTV; RX
Surgery	Laparoscopia; LPS; Laparotomia; LPT; Intervento chirurgico
Laboratory	CA-125; CA125; CEA; HE4; AFP; Markers; Marcatori
Therapy	NACT; Chemio; CHT; I LINEA; II LINEA; III LINEA
Biopsy / Histology	Biopsia; Istologico; EID

Table 1: Domain-specific acronyms and lexical variants used for event segmentation and tagging. All reported acronyms and keywords correspond to synonyms or abbreviations of the respective event type.

3.2.3 Regex + LLM + Recovery

Building on recent grounding techniques such as those proposed in MedSlice (Davis et al., 2025), this third strategy extends the previous cascaded regex+LLM pipeline by introducing an additional recovery procedure. This was necessary because, despite explicit instructions to extract segments verbatim, LLMs frequently introduce subtle rephrasings, summarize content, or omit punctuation. These "hallucinations of form" prevent direct string comparison and necessitate fuzzy alignment to map the LLM's semantic boundaries back to the original, legally-valid source text. After the initial regex-based date segmentation and the subsequent LLM refinement, this module verifies and reconstructs segments through the following steps:

1. **Fuzzy Matching.** Each LLM-generated segment is aligned to the original document using fuzzy string matching with the available Python library[1], which returns a similarity score s and the corresponding character offsets. Defining a threshold τ, if $s \geq \tau$, the segment is considered valid and the exact substring is extracted from the original text using the matched positions.

2. **Gap Filling.** Segments with $s < \tau$ are marked as invalid. When an invalid segment is located between two valid segments, it is replaced by directly recovering the original text spanning the gap between the end of the preceding valid segment and the start of the following one, ensuring complete coverage of the source document.

3. **Rollback.** If the proportion of the remaining invalid segments exceeds a predefined threshold ρ, the LLM-based segmentation is discarded entirely and the system reverts to the regex-only method.

Flow 3 in Figure 1 represents this approach. After a preliminary grid search over the alignment parameters, we selected the configuration that provided the best balance between alignment precision and coverage. Specifically, we set the similarity threshold to $\tau = 85\%$ and the Rollback ratio to $\rho = 0.5$, as this combination yielded the most stable and consistent performance across our experiments.

3.3 Event-Date Linking

After segmentation using one of the three previously described strategies, we introduce an additional processing step aimed at associating a temporal reference with each extracted event before performing event tagging.

[1] https://github.com/rapidfuzz/RapidFuzz

For segments obtained through date-based separation, the associated date is directly inherited from the date expression used as the segmentation boundary.

In contrast, for segments generated through the secondary step (which is acronym-based rules in the fully-regex strategy or LLM-based segmentation in the cascade approaches), the event remains temporarily without an associated date and subsequently handled during the post-processing phase.

3.4 Event Tagging

The date-linked segments are assigned a single event label through a pattern-matching procedure. Specifically, each segment is scanned for the presence of the acronyms listed in Table 1, which define the event categories. When multiple patterns are present within the same segment, the label corresponding to the earliest occurring pattern is selected. This strategy ensures a deterministic and consistent tagging process for identifying the key diagnostic events in the patient trajectory.

3.5 Post-Processing

After tagging, the segments undergo a cleaning and normalization procedure to ensure temporal consistency and structural coherence. First, dates are normalized by expanding two-digit years using a pivot rule (years ≤ 25 mapped to the 2000s and ≥ 26 to the 1900s) and assigning the first day of the month to month–year expressions. Segments without an explicit date inherit the temporal reference from the immediately preceding segment. When only month and year information is available and the preceding segment specifies a full date within the same month, the current date is shifted to one day after the previous event (this imputation rule was agreed upon with the clinical experts to maintain the chronological sequence of the patient's history). Malformed or truncated years are corrected by locating the nearest valid four-digit year and inferring the appropriate value based on month comparison. Finally, consecutive segments with the same event label and identical date are merged through text concatenation. This post-processing step mitigates over-segmentation errors, ensuring that a clinical event incorrectly divided into multiple fragments, is reconstructed as a single, semantically coherent unit.

4 Experiments

4.1 Input Data

The original dataset comprises approximately 1600 patients with more than 5000 clinical notes written in Italian language. From this larger corpus, we constructed a gold standard consisting of 343 text segments for evaluation, derived from reports belonging to 20 patients. During the gold standard annotation process, a text segment was strictly defined as an independent temporal event. The segments and their associated event tags were manually annotated by a clinical expert. Furthermore, to ensure an unbiased evaluation, all regex patterns and the acronym dictionary were initially designed and tuned on a separate development set of reports that do not overlap with the 20 patient reports used for testing.

An overview of the gold dataset is provided in Table 2, while Figure 2 illustrates the distribution of event types across reports.

Characteristic	Value
Number of Reports	20
Total Segments	343
Mean Segments per Patient	17.1

Table 2: Gold Standard dataset statistics.

4.2 Segmentation Setup

All experiments were conducted on a workstation equipped with an NVIDIA L40S-48C GPU (48 GB VRAM). For the hybrid segmentation strategies, LLM inference was executed using `llama.cpp`[2]. We evaluated two domain-specific medical language models: MedGemma (27B) (Sellergren et al., 2025) and MedicalQwen3 (14B) (Yang et al., 2025a), in particular we used `unsloth/medgemma-27b-it-GGUF` and `mradermacher/MedicalQwen3-Reasoning-14B-IT-i1-GGUF`. Both models were deployed in GGUF format using `Q6_K` quantization. Furthermore, evaluation on segmentation was assessed through the comparison among the gold and computed segments, using five complementary metrics. Table 3 reports such results averaged over segments for each experiment, while further details about the metrics are listed below:

Pk. Measures the probability that two points k tokens apart are incorrectly classified as belonging

[2]https://github.com/ggml-org/llama.cpp

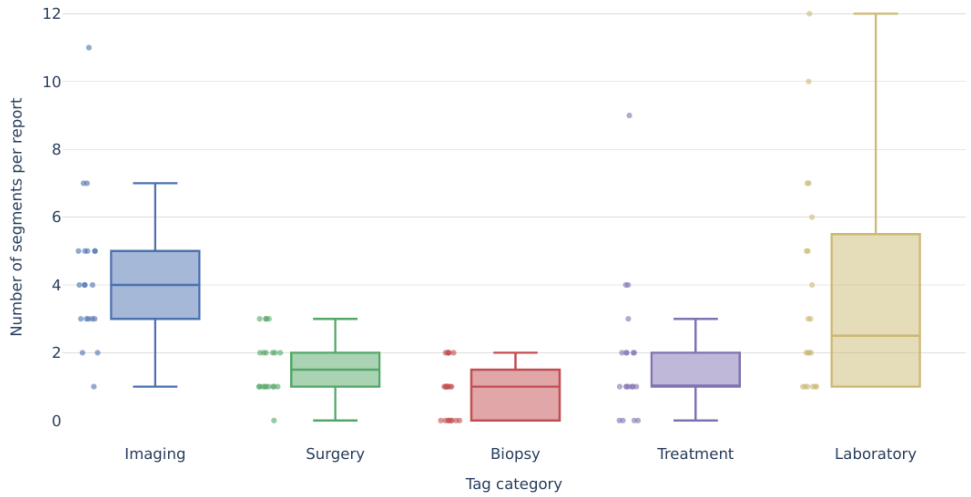


Figure 2: Distribution of segments per report, by tag event type.

to the same or different segments (lower is better) (Beeferman et al., 1999). For this metric, the pk function of the `SegEval` Python library was used.

WindowDiff. Counts boundary mismatches within a sliding window of size k, penalizing both missed and spurious boundaries (lower is better) (Pevzner and Hearst, 2002). For this metric, the `window_diff` function of the `SegEval` Python library was used (Fournier, 2013a).

Boundary Similarity. Computes an F1-style agreement between output and gold boundary positions (higher is better) (Fournier, 2013b). For this metric, the `boundary_similarity` function of the `SegEval` Python library was used (Fournier, 2013a).

Segment Count Accuracy. Quantifies structural agreement as $1 - |n_{\text{pred}} - n_{\text{gold}}|/n_{\text{gold}}$ (higher is better).

Text Overlap IoU. Measures the average Jaccard similarity between matched output and gold segments (higher is better).

Boundary-based metrics (Pk, WindowDiff, and Boundary Similarity) were computed at the *character level*. Predicted segments were converted into character-count ("mass") representations using the `SegEval` library (Fournier, 2013a). To account for minor formatting discrepancies introduced by LLMs (e.g., extra whitespaces), predicted masses were proportionally aligned to the gold-standard length through an `align_masses` procedure. For Pk and WindowDiff, the window size k was dynamically set to half the average gold segment length for each report. Conversely, Text Overlap IoU was computed at the *word level* using whitespace-tokenized segments and Jaccard similarity between greedily matched predicted and gold segments. Segment Count Accuracy and downstream F1 scores were computed globally across all reports.

4.3 Tagging Setup

Event classification of text segments was evaluated by comparing the gold and output segment tags using the F1 metric. The comparison relies on aligning gold-standard and predicted segments, which are not always perfectly matched: the number of output segments may differ from the gold standard, leading to over-segmentation or under-segmentation.

To address this, a fuzzy matching strategy was adopted. For each gold segment, the most semantically similar output segment is identified based on textual similarity. An output segment is considered a valid match only if its similarity with the corresponding gold segment exceeds an 80% threshold. This threshold is intentionally less strict than the one used in the recovery phase, as the goal here is to associate segments referring to the same underlying event type and enable a fair comparison during tagging evaluation, rather than to enforce precise boundary reconstruction.

Once matched, an output segment cannot be assigned to another gold segment. Unmatched gold segments (due to under-segmentation) are treated as False Negatives, while unmatched output seg-

Metric	Regex	Regex + MQ + Recovery	Regex + MG + Recovery	Regex + MQ	Regex + MG
Pk (↓)	**0.058 ± 0.080**	0.115 ± 0.094	0.076 ± 0.088	0.209 ± 0.125	0.149 ± 0.109
WindowDiff (↓)	**0.102 ± 0.096**	0.185 ± 0.110	0.125 ± 0.100	0.294 ± 0.139	0.225 ± 0.116
Boundary Sim. (↑)	**0.538 ± 0.353**	0.268 ± 0.265	0.459 ± 0.305	0.192 ± 0.249	0.288 ± 0.256
Seg. Count Acc. (↑)	**0.919 ± 0.091**	0.785 ± 0.209	0.855 ± 0.165	0.588 ± 0.311	0.679 ± 0.253
Text Overlap IoU (↑)	0.739 ± 0.248	**0.833 ± 0.127**	0.820 ± 0.196	0.806 ± 0.111	0.825 ± 0.149

Table 3: Segmentation metrics (mean ± std, $n = 20$) for the five setups. MQ = MedicalQwen3; MG = MedGemma. Arrows indicate whether lower (↓) or higher (↑) is better.

ments (due to over-segmentation or hallucinated boundaries) are treated as False Positives, ensuring that the reported F1 scores penalize segmentation errors.

Table 4 reports the results for each event type and overall across all experiments.

4.4 Results and Discussion

4.4.1 Segmentation Findings

Table 3 reports the segmentation performance across five configurations, where metrics are computed at the segment level and then averaged (mean ± std). Overall, the pure Regex strategy achieves the best results on most boundary-sensitive metrics, with the lowest Pk (0.058) and WindowDiff (0.102), and the highest Boundary Similarity (0.538) and Segment Count Accuracy (0.919), confirming the effectiveness of deterministic rules for precise boundary detection and segment preservation.

Among LLM-based approaches, MedGemma with Recovery shows the closest performance to Regex on structural metrics (Pk = 0.076; WindowDiff = 0.125; Seg. Count Acc. = 0.855), confirming the utility of Recovery in reducing boundary errors. Removing the Recovery mechanism leads to a clear degradation for both models (e.g., MedGemma Pk increases from 0.076 to 0.149), highlighting its stabilizing effect.

Interestingly, the best Text Overlap IoU is achieved by MedicalQwen3 with Recovery (0.833), slightly outperforming Regex (0.739). This suggests that, although LLM-based segmentation may not perfectly align with gold-standard boundaries, it captures semantic content.

The relatively low Boundary Similarity scores for LLM-based methods can be attributed to over- and undersegmentation effects. In particular, the introduction of spurious boundaries (oversegmentation) and occasional missed splits (undersegmentation) strongly penalize this F1-style boundary metric, even when Pk and WindowDiff indicate overall reasonable segmentation quality.

Finally, the relatively high standard deviations (especially for Boundary Similarity and Segment Count Accuracy) indicate notable inter-patient variability, likely due to differences in report structure and event density, which affect segmentation difficulty. Overall, Boundary Similarity is the most affected metric, showing high variability across all configurations, whereas Pk and WindowDiff remain more stable. The Regex approach exhibits the lowest standard deviations, likely because it was specifically engineered to handle the structural complexity of this dataset. However, when moving to different data sources, LLM-based techniques may offer greater generalizability.

4.4.2 Tagging Findings

Table 4 reports the F1 scores for event tagging under different segmentation strategies, both per category and overall. The results clearly show that segmentation quality directly impacts classification performance. In line with Table 3, the fully Regex-based segmentation (being the most accurate in boundary detection) also achieves the best overall F1 score (0.757), confirming that more precise segmentation leads to better downstream event classification.

At the category level, Regex outperforms the other methods in most cases (Imaging, Surgery, and Biopsy). However, the cascaded approaches with Recovery show competitive or superior results for Treatment (Regex + MedicalQwen3 + Recovery, 0.721) and Laboratory events (Regex + MedGemma + Recovery, 0.841), suggesting that for semantically richer or more variable segments, the LLM-enhanced strategy may better capture contextual cues.

Surgery achieves the highest F1 scores overall: it likely benefit from clearer structural markers and more explicit terminology, which facilitate both segmentation and tagging. In contrast, Biopsy shows the lowest performance across all config-

Event Tag	Regex	Regex + MQ + Recovery	Regex + MG + Recovery	Regex + MQ	Regex + MG
Imaging	**0.815**	0.749	0.812	0.682	0.709
Surgery	**0.923**	0.767	0.788	0.579	0.743
Biopsy	**0.605**	0.465	0.359	0.458	0.340
Treatment	0.697	**0.721**	0.675	0.628	0.659
Laboratory	0.759	0.794	**0.841**	0.817	0.835
Overall	**0.757**	0.668	0.720	0.623	0.652

Table 4: Per-category F1 scores for tag prediction. The Overall row reports micro-averaged F1 across all tags. MQ = MedicalQwen3; MG = MedGemma.

urations, likely because biopsy information is often embedded within surgical descriptions and not clearly separated, making both boundary detection and label assignment more challenging.

5 Conclusions

In this study, we investigated semantics-driven text segmentation as a core component for identifying temporal clinical events in oncological reports. By comparing a fully rule-based approach and a cascaded regex–LLM pipeline with and without a recovery mechanism, we demonstrated how segmentation quality directly impacts downstream event type identification.

The results show that regex-based segmentation ensures superior boundary precision and structural stability, while LLM-enhanced approaches better preserve semantic coherence within segments. These findings support the view that segmentation should be treated as a preprocessing step in order to make reports more easily processable for downstream analysis, but it also consists of a modelling task that enables reliable reconstruction of the patient's clinical trajectory.

As future work, we aim to extend this study by scaling up the evaluation dataset and involving multiple annotators to improve the reliability of the gold standard. We also plan to validate the framework on multiple external, heterogeneous datasets to further assess generalization. Methodologically, future iterations will explore the LLM-only segmentation on the full-report to better isolate the contribution of the recovery mechanism, and conduct an ablation study to decouple the impact of gap-filling and global rollback. Finally, we aim to replace the rule-based event tagger with a transformer-based classifier to evaluate downstream performance.

6 Limitations

While the results are promising, several directions for further improvement emerge. Expanding the evaluation dataset beyond the current 343 segments would enable a more robust assessment of generalization, and involving multiple annotators would strengthen the reliability of the gold standard. From a modeling perspective, evaluating LLMs in full precision (without quantization) could better capture their intrinsic capabilities, while a systematic optimization of the recovery mechanism may further enhance performance. Finally, extending event tagging beyond regex-based rules toward transformer-based approaches represents a natural step toward a more flexible and fully data-driven pipeline.

Ethics Statement

The use of data for this study has been implemented in full compliance with ethics and GDPR requirements. Specifically, data usage has been approved by the Ethics Committee of our hospital to conduct the presented research and the de-identification of sensitive data has been performed. Approval protocol number from the relevant Ethics Committee can be provided on request.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *Preprint*, arXiv:2205.12689.

Joe Barrow, Raj Patel, Misha Kharkovski, Ben Davies, and Ryan Schmitt. 2025. Safepassage: High-fidelity information extraction with black box llms. *Preprint*, arXiv:2510.00276.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Joshua Davis, Thomas Sounack, Kate Sciacca, Jessie M Brain, Brigitte N Durieux, Nicole D Agaronnik, and Charlotta Lindvall. 2025. Medslice: Fine-tuned large language models for secure clinical note sectioning. *Preprint*, arXiv:2501.14105.

Joshua Denny, Anderson Spickard, Kevin Johnson, Neeraja Peterson, Josh Peterson, and Randolph Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association : JAMIA*, 16:806–15.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. *Preprint*, arXiv:2106.07799.

Chris Fournier. 2013a. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, page to appear, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Fournier. 2013b. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941.

Akshay Goel. 2026. LangExtract.

Aman Jaiswal and Evangelos Milios. 2023. Breaking the token barrier: Chunking and convolution for efficient long text classification with bert. *arXiv preprint arXiv:2310.20558*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ricardo Kleinlein, Kathryn Gray, David Bates, and Vesela Kovacheva. 2025. Spell-llms: A scalable and privacy-compliant nlp pipeline using locally hosted large language models for clinical information extraction.

Livia Lilli, Carlotta Masciocchi, Antonio Marchetti, Giovanni Arcuri, and Stefano Patarnello. 2025a. Prompting large language models for italian clinical reports: A benchmark study. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 190–200.

Livia Lilli, Stefano Patarnello, Nikola Capocchiano, Carlotta Masciocchi, and Mario Santoro. 2025b. Improving clinical report classification with sentence boundary detection. pages 190–196.

Livia Lilli, Andrea Rosati, Giovanni Paolo Tobia, Massimo Criscione, Federica Tomassini, Chiara Dachena, Alice Luraschi, Chiara Cantarini, Carolina De Maria, Luigi Congedo, Massimo Bernaschi, Stefano Patarnello, and Anna Fagotti. 2026. Prompt-orchestrated large language models for clinical information extraction. *Research Square*. Preprint.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Nipun Sadvilkar and Mark Neumann. 2020. Pysbd: Pragmatic sentence boundary disambiguation. *arXiv preprint arXiv:2010.09657*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, 20.

Risha Surana, Adrian Law, Sunwoo Kim, Rishab Sridhar, Angxiao Han, and Peiyu Hong. 2025. Cnsight: Evaluation of clinical note segmentation tools. *Preprint*, arXiv:2512.22795.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Shin, Kaleb Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony Costa, Mona Flores, Ying

Zhang, Tanja Magoc, Christopher Harle, Gloria Lipori, Duane Mitchell, William Hogan, Elizabeth Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5.

Zongxin Yang, Hongyi Yuan, Raheel Sayeed, Amelia Li Min Tan, Enci Cai, Mohammed Moro, Xiudi Li, Huaiyuan Ying, Nicholas Brown, Griffin Weber, Sheng Yu, Isaac Kohane, and Tianxi Cai. 2025b. Clines: Clinical llm-based information extraction and structuring agent. *medRxiv*.

Operation-Mechanism Alignment for Reliable Large Language Model Reasoning over Electronic Health Records

Guanyu Tao[1], **Siyao Wang**[1,2], **Yong Xue**[1], **Ashwani Tanwar**[1]
Yuting Ji[1], Kai Sun[1,2], Monica Mok[1], Marzana Chowdhury[1]
Deepa Gupta[1], Ashok Gupta[1], Jingqing Zhang[1*], Vibhor Gupta[1], Yike Guo[1,3]
[1]Pangaea Data Limited, UK and USA
[2]Data Science Institute, Imperial College London, London, UK
[3]Hong Kong University of Science and Technology, Hong Kong SAR, China

Abstract

Clinical reasoning over electronic health records (EHRs) involves heterogeneous operations, including text interpretation, numerical computation, temporal filtering, and guideline-based aggregation. However, many existing LLM-based approaches still cast these heterogeneous operations as a single end-to-end generation process, obscuring their different reliability requirements and making intermediate failures difficult to inspect. We therefore propose a framework based on operation-mechanism alignment that represents clinical reasoning as a directed acyclic graph of typed operations, where each node is assigned to the execution mechanism best suited to its reliability requirements. The framework also preserves structured evidence provenance for intermediate results. Across six clinician-annotated binary decision tasks, the framework outperforms direct prompting, single-step retrieval-augmented prompting, and chain-of-thought baselines, supporting operation-mechanism alignment as a practical design principle for reliable clinical reasoning over EHRs.

1 Introduction

Clinical reasoning over electronic health records (EHRs) often requires several different types of operations within the same task. For example, staging chronic kidney disease requires interpreting free-text nephrology notes to assess proteinuria, computing eGFR from a validated formula, selecting laboratory values within clinically relevant time windows, and combining the findings under guideline-defined logic. These steps have different reliability requirements. Numerical computation and temporal filtering require exact execution, because an arithmetic or date-filtering error can change the assigned disease stage. Proteinuria assessment, by comparison, depends on flexible interpretation of heterogeneous clinical language. Treating these steps as a single end-to-end generation problem obscures these operational differences and makes failures difficult to trace, inspect, or correct.

Common prompting-based approaches still cast clinical reasoning as a single inference pass. The model is prompted over either the full patient record or a retrieved subset of documents. Although retrieval-augmented generation can improve evidence relevance, the same generation step must still select evidence, perform calculations, apply clinical rules, and produce the final decision. This creates two practical problems. First, errors in intermediate operations, particularly numerical computation and temporal filtering, can propagate silently to the final answer. Second, the output is difficult to inspect because intermediate decisions and evidence assignments are not explicitly represented.

We argue that clinical reasoning should be decomposed into typed operations, including text interpretation, numerical computation, temporal filtering, and criteria aggregation. Different operation types should not be handled by the same generation step. Language interpretation can be handled by LLM-based evaluators, numerical and temporal operations by deterministic tools, and criteria aggregation by explicit logical operators. We call this principle *operation-mechanism alignment*. We implement this principle by representing clinical reasoning as a directed acyclic graph, where each node corresponds to a typed operation and each intermediate result carries structured evidence provenance. This makes the reasoning process inspectable at each step, rather than hidden behind a single generation call.

We evaluate the framework on clinician-annotated clinical reasoning tasks and compare it against direct prompting, single-step retrieval-augmented prompting, and chain-of-thought prompting baselines. Our contributions are:

*Correspondence to jzhang@pangaeadata.ai

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 605–616
July 3-4, 2026 ©2026 Association for Computational Linguistics

(1) we formalize clinical reasoning over EHRs as a graph of typed operations with explicit operation-mechanism alignment, rather than single-pass generation; (2) we present a framework that routes evaluation, computation, and aggregation to distinct execution mechanisms while maintaining evidence provenance throughout; and (3) we provide an evaluation showing the benefits of this design over standard prompting and retrieval-based baselines.

We evaluate the framework on clinician-annotated clinical reasoning tasks and compare it with baselines including direct prompting, single-step retrieval-augmented prompting, and chain-of-thought prompting. Our contributions are: (1) we formalize EHR-based clinical reasoning as a graph of typed operations with explicit alignment between operation types and execution mechanisms; (2) we implement an executable framework that routes language evaluation, deterministic computation, temporal filtering, and logical aggregation to distinct mechanisms while preserving evidence provenance; and (3) we provide an evaluation showing the benefits of this structured design over standard prompting and retrieval-based baselines.

2 Related Work

Recent benchmarks study clinical reasoning with large language models in patient-specific EHR settings, requiring recovery of clinical state from records rather than decontextualized questions or pre-structured cases. EHRNoteQA and ER-REASON operationalize this setting through multi-note questions, emergency-department workflow tasks, longitudinal heterogeneous notes, and clinician- or physician-authored supervision (Kweon et al., 2024; Mehandru et al., 2025). DR.BENCH and MedHELM further show that clinical reasoning spans tasks with different operational demands, motivating methods beyond monolithic prompt-to-answer formulations, while related application studies apply LLMs directly to gold-labeled EHRs in largely one-pass prompting settings (Gao et al., 2023; Bedi et al., 2025; Zhang et al., 2024).

Retrieval-augmented generation (RAG) is widely used to improve grounding in medical question answering and clinical reasoning over EHRs. MIRAGE/MedRAG shows that RAG can outperform chain-of-thought prompting while exposing retrieval-specific issues such as context position effects (Xiong et al., 2024b). i-MedRAG and Self-BioRAG explore multi-round evidence acquisition and self-critique for complex questions and retrieval noise (Xiong et al., 2024a). In EHR settings, ExpRAG uses case-based retrieval from other patients' discharge reports, while EHR-RAG combines temporally and event-aware retrieval with iterative refinement for long-horizon prediction (Ou et al., 2025; Cao et al., 2026). However, these approaches still typically execute clinical reasoning over EHRs within a single generative mechanism, even when it involves temporal filtering, numeric aggregation, and guideline-like decision logic.

Temporal and longitudinal reasoning remains challenging for LLMs in clinical reasoning over EHRs. TIMER introduces a temporal benchmark and instruction-tuning method grounded in time-indexed patient-record segments, suggesting that time-awareness is not captured by standard instruction tuning (Cui et al., 2025). Long-context clinical summarization and prediction studies similarly show that even with longer context windows, RAG variants, and chain-of-thought prompting, models struggle to maintain chronological coherence over extended patient trajectories (Kruse et al., 2025). These findings suggest that some EHR reasoning subproblems, especially temporal filtering and longitudinal aggregation, may require explicit operators such as time-window queries rather than free-form generation.

Graph-structured orchestration and agentic tool use provide another adjacent direction. MedAgentBench offers a Fast Healthcare Interoperability Resources (FHIR)-compliant environment for benchmarking medical agents on patient-specific record workflows (Jiang et al.). DSPy provides graph- or pipeline-based composition patterns, ReAct and Toolformer study tool invocation, and Graph-of-Thought extends linear chains to graph-structured intermediate thoughts (Khattab et al., 2023; Yao et al., 2022; Schick et al., 2023; Besta et al., 2024). However, these systems mainly use graphs to organize reasoning steps or tool calls, rather than define operation semantics or use operation types to determine execution mechanisms.

Another concern in clinical reasoning over EHRs with LLMs is how outputs are grounded, verified, and linked to patient-record evidence. VeriFact studies patient-record-grounded verification of generated clinical narratives, while FactEHR shows that clinical fact decomposition varies across LLMs

and note types (Chung et al., 2025; Munnangi et al., 2024). Related evaluation work validates LLM-as-a-judge assessment of multi-document EHR summaries aligned with clinician rubrics (Croxford et al., 2025). Standards such as CDS Hooks and CPG-on-FHIR show how guideline logic and evidence links can be made computable and integrated with EHR workflows. However, much of this literature attaches evidence to final outputs through verification, evaluation, or post hoc citation, rather than preserving provenance throughout intermediate reasoning steps.

In summary, existing work has improved clinical reasoning over EHRs through prompting, retrieval, tool use, and output verification. To the best of our knowledge, this is the first work to model clinical reasoning over EHRs as heterogeneous, typed operations routed to mechanisms aligned with reliability requirements, while preserving structured provenance across intermediate results.

3 Approach

Our central design principle is *operation-mechanism alignment*: clinical reasoning over EHRs should be modeled as a composition of heterogeneous operations with different reliability requirements, rather than as a monolithic generation problem. In our framework, the execution mechanism depends on the operation type. Language-mediated clinical judgment is handled by LLM-based modules; numeric, temporal, and unit-sensitive computation is handled by deterministic tools; and criteria aggregation is handled by explicit logical operators. This hybrid neural-symbolic design preserves interpretive flexibility while reducing avoidable errors in exactly computable operations.

Throughout this section, we use a **chronic kidney disease (CKD) assessment** task as a running example. The clinical question is binary: *does this patient have advanced-stage CKD with rapid kidney function decline?* Clinically, this decision is driven primarily by the estimated glomerular filtration rate (eGFR): advanced stage corresponds to eGFR below a threshold, and rapid decline to the eGFR decline rate over time. Answering it therefore requires interpreting clinical notes and laboratory reports to determine whether the patient has persistent albuminuria as a supporting criterion (evaluation), computing current and historical eGFR values from laboratory measurements and estimating the annual decline rate as the central quantitative driver (computation), and combining these results under clinical guideline logic to produce a yes-or-no decision (aggregation). Relevant clinical details are introduced as each operation type is discussed.

3.1 Problem Formulation

We define a clinical reasoning task template as $T = (Q, C)$, where Q is a task query describing the clinical question to be answered and C is a set of clinical criteria and decision constraints relevant to the task. Given a patient record D, the goal is to predict a binary label $\hat{y} \in \{0, 1\}$.

In the CKD example, Q asks whether the patient has advanced-stage CKD with rapid decline. C includes an eGFR threshold for advanced stage, an annual eGFR decline threshold for rapid progression, albuminuria severity rules, and a persistence requirement ($\geq$3 months). D comprises laboratory results, clinical notes, and demographics.

To solve (T, D), we instantiate a directed acyclic graph $G = (V, E, \tau)$, where each node $v_i \in V$ is associated with a typed operation $\tau(v_i) \in \{\text{EVALUATE}, \text{COMPUTE}, \text{AGGREGATE}\}$ and edges encode dependency constraints. Each node is executed over node-specific evidence e_i derived from D for that subproblem, together with any required upstream outputs $\{z_j : (v_j, v_i) \in E\}$. For some nodes—particularly aggregation nodes that only compose upstream results—e_i may be empty and execution depends entirely on upstream outputs. The node output $z_i = (r_i, \mathcal{E}_i)$ is a structured intermediate result consisting of an operation result r_i and supporting evidence references $\mathcal{E}_i$.

- **Evaluation nodes** determine whether a clinical condition is satisfied using language-mediated reasoning over retrieved evidence; in the running example, this includes assessing persistent albuminuria from notes and laboratory reports.
- **Computation nodes** invoke deterministic operators for numeric, temporal, or unit-sensitive operations; in the running example, these compute current and historical eGFR values and the annualized decline rate.
- **Aggregation nodes** combine upstream outputs into higher-level criterion decisions or the final prediction $\hat{y}$; in the running example, eGFR-based staging serves as the pri-

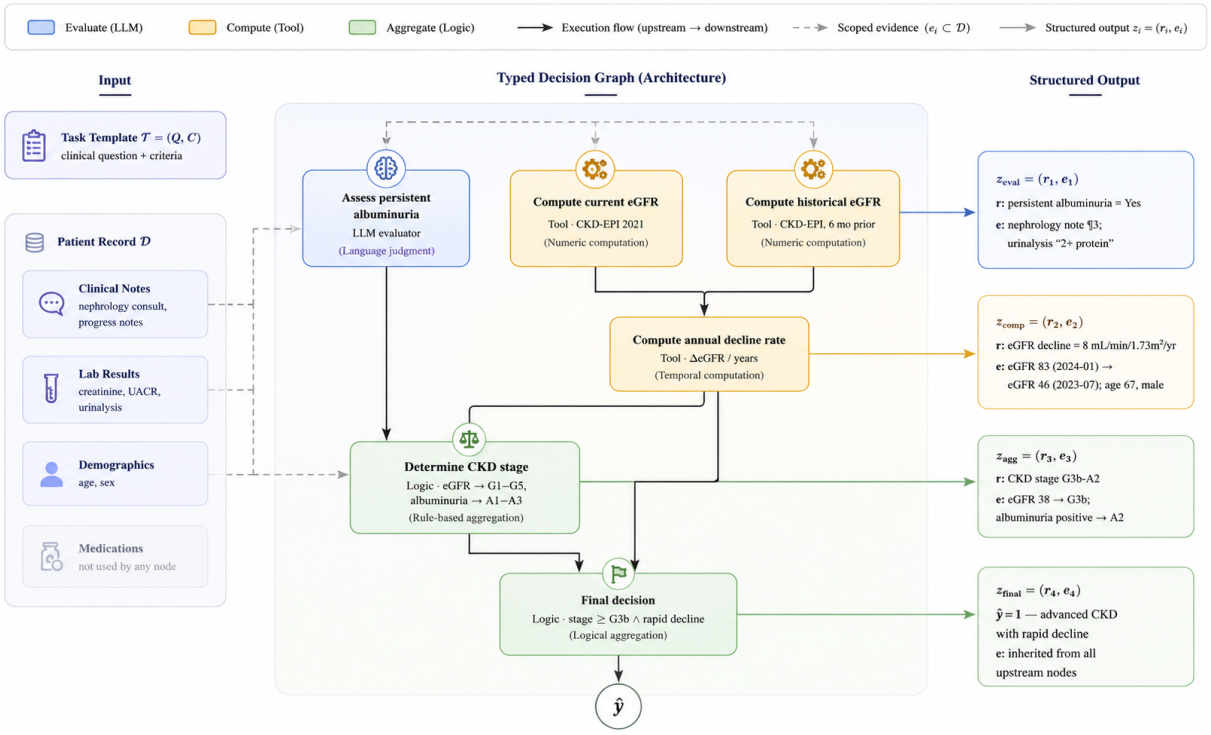


Figure 1: Overview of our framework. A clinical reasoning task template and patient record are instantiated as a typed decision graph. Each node is routed to the execution mechanism aligned with its operation type: LLM-based evaluators for language-mediated judgment, deterministic tools for numeric and temporal computation, and explicit logical operators for aggregation. Node outputs are structured and evidence-grounded, enabling provenance-preserving execution traces. Dashed arrows indicate scoped evidence: different nodes operate over different subsets of the patient record, rather than a single shared context.

mary backbone, combined with the albuminuria category and the progression criterion under guideline logic.

Figure 1 illustrates this decomposition for the CKD running example, showing how different patient record sources provide scoped evidence to specific nodes, how nodes are routed to distinct execution mechanisms, and how each node output carries structured evidence references.

3.2 Task Decomposition and Execution

A key property of the graph is that each node operates on a *scoped* subproblem with its own evidence requirements, rather than on a single shared context. This contrasts with retrieval-augmented approaches, where one context bundle must support evidence identification, calculation, temporal reasoning, and final decision making at once. In our framework, evidence is constructed locally: proteinuria evaluation uses clinical notes and urinalysis reports, whereas eGFR computation uses structured laboratory observations within defined temporal windows.

Edges in G encode explicit data dependencies. The eGFR decline computation depends on both the current and historical eGFR computations; the final aggregation depends on the eGFR-based staging, the decline rate, and the albuminuria evaluation. This dependency structure determines a partial execution order: independent nodes—such as the albuminuria evaluation and the two eGFR computations—can execute in parallel, while dependent nodes wait for upstream results.

The graph G is instantiated from the task template T before execution. Each criterion in C maps to one or more nodes in V, and dependencies are derived from the logical relationships among criteria. This turns a declarative task specification into an executable plan, separating *what* must be decided from *how* each subproblem is resolved.

3.3 Operation-Mechanism Alignment

The three node types are not merely a descriptive taxonomy; they determine which execution mechanism is invoked. This section describes how each mechanism operates and why the routing matters.

Evaluation: LLM-based clinical judgment. Evaluation nodes handle operations that require interpreting diverse and often ambiguous clinical language. In the running example, determining whether a patient has persistent albuminuria may involve reasoning over a urinalysis showing "2+ protein," a nephrology consult noting "persistent albuminuria documented over the past 4 months, now in the moderately increased range," and structured uACR (urine albumin-to-creatinine ratio) measurements—all expressed differently and scattered across the record. These judgments resist rule-based encoding because the space of possible phrasings is open-ended. LLM-based reasoning is therefore the appropriate mechanism for evaluation nodes: it provides the flexibility to interpret heterogeneous clinical text, at the cost of being non-deterministic.

Computation: deterministic tools. Computation nodes handle operations where correctness is exactly verifiable and LLM generation introduces unnecessary risk. In the running example, computing the patient's estimated glomerular filtration rate (eGFR) requires applying a validated clinical formula (CKD-EPI 2021) to serum creatinine, age, and sex. A small numeric error—an eGFR of 46 versus 38—can shift the patient between adjacent disease stages, directly changing the clinical decision. The same concern applies to unit conversion (creatinine reported in mg/dL versus μmol/L), temporal filtering (selecting laboratory values within a specific time window), and rate-of-change computation (pairing values across time points and annualizing the difference). These operations are delegated to validated deterministic tools: clinical calculators for formula evaluation, unit-aware conversion routines, temporal window filters, and arithmetic operators for derived quantities.

Aggregation: explicit logical composition. Aggregation nodes combine upstream results under predefined clinical logic rather than open-ended synthesis. In the CKD example, the final staging decision follows clinical guidelines: map eGFR to a disease stage, combine it with the albuminuria category, and check the progression criterion. This composition is a deterministic function of its inputs and should not be subject to LLM variability. Aggregation therefore uses explicit logical operators—boolean combinations, threshold comparisons, and lookup tables—to produce the final prediction $\hat{y}$.

Why alignment matters. The key claim is not that any one mechanism is universally superior, but that misalignment between operation type and execution mechanism introduces avoidable errors. Delegating eGFR computation to an LLM risks arithmetic mistakes in a context where precision determines the clinical outcome. Conversely, attempting to encode proteinuria assessment as a rule-based system would require enumerating an impractical number of textual patterns. Operation-mechanism alignment is the principle that each node should be executed by the mechanism whose failure modes are least harmful for that operation's reliability requirements.

3.4 Evidence-Grounded Output

Every node output $z_i = (r_i, \mathcal{E}_i)$ includes structured evidence references $\mathcal{E}_i$ that link the result back to specific records in the patient record D. These references are not post-hoc citations but are constructed during execution: evaluation nodes record which passages informed their judgment, computation nodes record which laboratory values and timestamps were used, and aggregation nodes inherit the evidence from all upstream nodes they compose.

This design serves two purposes. First, it supports clinical auditability: a reviewer can trace any component of the final decision back to its source evidence without re-running the system. Second, it makes unsupported conclusions structurally detectable—if a node produces a result without corresponding evidence references, this is visible in the output structure rather than hidden in a free-text rationale.

In the running example, the final decision that a patient has advanced-stage CKD with rapid decline is accompanied by the specific laboratory values and their timestamps used to compute both eGFR values, the clinical note passages that supported the albuminuria assessment, and the guideline rule that combined these into the staging decision. Each component is individually inspectable.

4 Experiments

4.1 Data and Tasks

We evaluate the framework on de-identified EHR data from selected subsets of MIMIC-III and MIMIC-IV, together with a de-identified partner-hospital dataset (from Cone Health). The evaluation covers six binary clinical decision tasks under

a shared annotation protocol.

COPD exacerbation history. The first task asks whether a patient has a confirmed recent COPD exacerbation history, defined by recurrent outpatient exacerbation treatment, at least one inpatient hospitalization for acute COPD exacerbation, or repeated prescription courses of systemic corticosteroids or antibiotics for exacerbation within the past year. The task is evaluated on two cohorts under the same task definition: 151 MIMIC-III patients (108 positive, 43 negative), and 100 patients from the de-identified Cone Health dataset (18 positive, 82 negative).

Asthma exacerbation risk. The second task determines whether a patient with confirmed asthma is at risk of exacerbation or asthma-related death under the Global Initiative for Asthma (GINA) 2024 criteria (Global Initiative for Asthma, 2024), including severe exacerbation history, poor symptom control, inadequate ICS treatment, high SABA use, low FEV_1, relevant comorbidities, or prior near-fatal events. The cohort comprises 50 MIMIC-IV patients (46 positive, 4 negative).

Cancer cachexia screening (Fearon). The third task applies the cachexia criteria of Fearon et al. (Fearon et al., 2011), requiring an underlying cancer diagnosis together with weight loss $>5\%$ over 6 months, or BMI <20 kg/m^2 with weight loss $>2\%$, or sarcopenia with weight loss $>2\%$. The cohort comprises 50 patients (15 positive, 35 negative) from MIMIC-III and IV.

Cancer cachexia screening (Evans). The fourth task applies the cachexia criteria of Evans et al. (Evans et al., 2008), requiring an underlying chronic disease including cancer, involuntary weight loss $\geq 5\%$ in the past 12 months, and at least three of five additional features: decreased muscle strength, anorexia, low fat-free mass index, abnormal biochemistry, or decreased functional status. The cohort comprises 50 patients (8 positive, 42 negative) from MIMIC-III and IV.

Confirmed CKD diagnosis. The fifth task asks whether the patient has an explicit documented diagnosis of chronic kidney disease (CKD) in the clinical record. The cohort comprises 100 patients from the de-identified Cone Health dataset (25 positive, 75 negative).

At-risk CKD identification. The sixth task asks whether a patient is at risk of CKD but does not yet have a confirmed diagnosis. Positive cases require at least one recognized CKD risk factor or non-chronic kidney abnormality, including hypertension, diabetes mellitus, cardiovascular disease, prior AKI/AKD, reduced eGFR, elevated albuminuria, urine sediment abnormalities, hematuria, histology abnormalities, tubular or electrolyte disorders, or structural kidney abnormalities on imaging, while excluding patients who meet confirmed CKD criteria. The cohort comprises 100 patients from the same Cone Health dataset (42 positive, 58 negative).

Gold labels are created by clinician annotators on curated subsets using task-specific criteria. For each case, annotators provide the expected binary label and, when applicable, supporting evidence from the record. Each case is reviewed by two clinician annotators independently and any disagreement is resolved by a third clinician annotator or with consensus.

All methods are evaluated using GPT-5.4 as the primary model; direct-prompting sensitivity to model choice is additionally assessed with GPT-5.2 and Claude Opus 4.6 (see Appendix A).

4.2 Baselines

We compare the proposed framework against three baseline settings that use the same underlying patient records and task definitions, but differ in how evidence is constructed and how reasoning is organized:

- **Direct prompting**: The patient record is provided to the model in a single pass with a task instruction, without explicit task decomposition, intermediate structure, or tool use. This baseline tests whether a general-purpose prompt over the available patient context is sufficient for the binary decision.

- **Single-step RAG**: Task-relevant evidence is retrieved once using the task query, and the retrieved context is then provided to the model for end-to-end reasoning in a single generation step. This baseline isolates the value of retrieval while still requiring one model call to perform evidence interpretation, calculation, temporal reasoning, and final decision making jointly.

- **Chain-of-thought (CoT) prompting**: The model receives the same retrieved context as

in the single-step RAG baseline, but is explicitly prompted to reason step-by-step before producing the final label. This baseline tests whether prompting for intermediate natural-language reasoning alone is sufficient, without explicit task-scoped execution or deterministic tool support.

For fairness, all methods use the same underlying task definitions, model set, and binary output schema. When the full patient record exceeds the model context budget, the direct-prompting baseline uses a shared record-linearization pipeline with budget-matched truncation, while retrieval-based baselines and our framework use the same task description and retriever configuration wherever applicable. These baselines are therefore designed to separate the effects of retrieval, explicit reasoning structure, and tool-supported execution rather than by differences in model access or output format.

4.3 Metrics

Because our gold labels are binary clinical decisions, we use sensitivity, specificity, precision, and F1 as the primary task metrics. Accuracy is included as a supplementary metric.

4.4 Results

Table 1 reports the main comparison across reasoning strategies using GPT-5.4 on all six tasks.

On COPD exacerbation, the framework achieves the best F1 and accuracy on both cohorts. On MIMIC-III, it reaches F1 = .893 and accuracy = .861, with a marked increase in specificity (.977 versus .349 for direct prompting and .186 for both retrieval-based baselines). The same pattern holds on Cone Health, where the framework attains F1 = .882 and specificity = .988, while baseline specificities range from .780 to .866; the four methods identify the same set of 15 true positives but the framework yields only 1 false positive versus 11–18 for the baselines. The consistent specificity advantage across two independent datasets is consistent with baseline methods over-predicting positive cases when temporal and diagnostic attribution must be resolved precisely, whereas deterministic temporal filtering and explicit diagnostic-attribution logic are associated with fewer false positives. On asthma exacerbation risk, the framework again achieves the best F1 (.957) and accuracy (.920) and is the only method with nonzero specificity (.500, 2/4). This specificity estimate should be interpreted cautiously because the task contains only four negative cases.

On cancer cachexia screening, the framework achieves the highest F1 and accuracy for both Fearon and Evans criteria. For the Fearon criteria, it attains F1 = .839 and accuracy = .900, while chain-of-thought prompting performs worst (F1 = .581). This result is consistent with the difficulty of applying quantitative thresholds, such as weight loss, BMI, and sarcopenia indices, through free-text step-by-step reasoning alone. For the more stringent Evans criteria, the framework reaches sensitivity = 1.00 and F1 = .727, compared with the best baseline F1 of .480, consistent with the difficulty of aggregating multiple concurrent conditions in a single-pass setup.

On the CKD tasks evaluated on Cone Health data, the framework achieves the highest F1 and accuracy on both tasks. For confirmed CKD diagnosis, it reaches F1 = .909 with perfect sensitivity (1.000), recovering all 25 confirmed cases; chain-of-thought and single-step RAG both reach .960 sensitivity but with one missed case each, while direct prompting lags further (F1 = .792, sensitivity = .760). Specificity is comparable across the four methods (.920–.947) on this task, so the framework's advantage concentrates on sensitivity rather than specificity. For at-risk CKD identification, the framework attains F1 = .900 and accuracy = .920, with the highest precision (.947) and specificity (.966); chain-of-thought is competitive on sensitivity (.881 versus .857 for the framework), but produces more false positives. Together, these results suggest that detailed multi-criteria CKD definitions can challenge single-pass reasoning, while decomposition into independent criteria evaluations is associated with stronger performance in these tasks.

Model sensitivity. As a secondary check, we repeated direct prompting with GPT-5.2 and Claude Opus 4.6. Performance varied by task and model, but model substitution did not eliminate the limitations of single-pass prompting; full results are provided in Appendix A. We therefore treat model choice as an important but orthogonal factor.

4.5 Discussion

The results reveal several consistent patterns across tasks.

Specificity as the differentiator. Across the six tasks, the proposed framework consistently achieves the highest or near-highest specificity.

Task	Method	Sens.	Spec.	Prec.	F1	Acc.
COPD Exacerbation (MIMIC-III)	Direct prompting	**.843**	.349	.765	.802	.702
	Single-step RAG	**.843**	.186	.722	.778	.656
	Chain-of-thought	.806	.186	.713	.757	.629
	Ours	.815	**.977**	**.989**	**.893**	**.861**
COPD Exacerbation (Cone Health)	Direct prompting	.778	.780	.438	.560	.780
	Single-step RAG	**.833**	.866	.577	.682	.860
	Chain-of-thought	**.833**	.829	.517	.638	.830
	Ours	**.833**	**.988**	**.938**	**.882**	**.960**
Asthma Risk†	Direct prompting	.913	.000	.913	.913	.840
	Single-step RAG	**.935**	.000	.915	.925	.860
	Chain-of-thought	.913	.000	.913	.913	.840
	Ours	.957	**.500**	**.957**	**.957**	**.920**
Cachexia (Fearon)	Direct prompting	.800	.743	.571	.667	.760
	Single-step RAG	.733	.829	.647	.688	.800
	Chain-of-thought	.600	.800	.563	.581	.740
	Ours	**.867**	**.914**	**.813**	**.839**	**.900**
Cachexia (Evans)	Direct prompting	.875	.452	.233	.368	.520
	Single-step RAG	.875	.619	.304	.452	.660
	Chain-of-thought	.750	.738	.353	.480	.740
	Ours	**1.00**	**.857**	**.571**	**.727**	**.880**
Confirmed CKD	Direct prompting	.760	**.947**	.826	.792	.900
	Single-step RAG	.960	.933	.828	.889	.940
	Chain-of-thought	.960	.920	.800	.873	.930
	Ours	**1.000**	.933	**.833**	**.909**	**.950**
At-Risk CKD	Direct prompting	.833	.862	.814	.824	.850
	Single-step RAG	.833	.948	.921	.875	.900
	Chain-of-thought	**.881**	.931	.902	.892	.910
	Ours	.857	**.966**	**.947**	**.900**	**.920**

Table 1: Main results across reasoning strategies (GPT-5.4). Bold indicates the best result per metric per task. Sens. = sensitivity, Spec. = specificity, Prec. = precision, Acc. = accuracy. †High class imbalance (46/4); specificity estimates are unstable due to very few negative cases.

This is most pronounced in COPD exacerbation across both cohorts (.977 vs. .349 for direct prompting on MIMIC-III; .988 vs. .780–.866 on Cone Health), cachexia under Evans criteria (.857 vs. .452), and asthma exacerbation risk, where all three baselines achieve zero specificity while the framework is the only method that correctly rejects any negative cases. On at-risk CKD, the framework reaches .966 specificity versus .862–.948 for baselines, with .947 precision versus .814–.921, suggesting it is more conservative in rejecting borderline cases that satisfy only one or two soft risk factors. On confirmed CKD, specificity differences across methods are small (.920–.947); the framework's lead instead concentrates on sensitivity (1.000 versus .760–.960), recovering all confirmed cases. The framework's explicit criteria decomposition and deterministic aggregation logic enforce the full set of diagnostic requirements (e.g., temporal windows, primary-diagnosis attribution, multi-condition conjunctions), preventing the model from short-circuiting to an affirmative answer when only partial evidence is present.

Structured execution outperforms unstructured reasoning. Chain-of-thought (CoT) prompting, which encourages step-by-step reasoning in natural language, does not consistently improve over direct prompting and in some cases degrades performance. On cachexia (Fearon), CoT achieves the lowest F1 of any method (.581 vs. .667 for direct prompting). This suggests that prompting the model to "think step by step" does not substitute for routing quantitative operations to deterministic tools. When intermediate steps involve numeric thresholds, weight-loss calculations, or temporal window enforcement, natural-language reasoning introduces opportunities for error that structured execution avoids.

Task complexity modulates framework advantage. The framework's advantage is largest on tasks requiring multi-step quantitative criteria and smallest on simple diagnostic lookup. These tasks share a common structure: evaluating multiple in-

dependent clinical criteria with their own evidence, chronicity, and exclusion rules. Providing such detailed criteria as a single instruction degrades baseline performance, which the framework avoids by decomposing the specification into independent, single-feature evaluations.

Relation to tool-using agents. Autonomous tool-using agents (e.g., ReAct-style systems (Yao et al., 2022) or Toolformer (Schick et al., 2023)) address a distinct problem: discovering task decomposition at inference time when the decomposition is unknown. In our setting the decomposition is known—each task graph is derived from a validated clinical guideline (e.g., GOLD 2025, KDIGO 2024). Letting an agent re-discover this structure at inference replaces a vetted plan with one assembled on the fly, sacrificing auditability without addressing how each step should be executed. The latter question—operation-mechanism alignment—is orthogonal to tool use: even a ReAct agent with calculator access must still decide when to invoke the calculator versus reason in natural language, which is the same alignment decision our framework resolves ahead of time. Our contribution therefore complements rather than substitutes for autonomous tool use.

5 Conclusion

We presented a framework that structures clinical reasoning over EHRs as a directed acyclic graph of typed operations, where each node is routed to the execution mechanism—LLM-based evaluation, deterministic computation, or explicit logical aggregation—best suited to its reliability requirements. Every intermediate result carries structured evidence provenance, making the reasoning process inspectable at each step.

Experiments on six clinician-annotated tasks across four disease areas show that this design consistently improves specificity over direct prompting, retrieval-augmented, and chain-of-thought baselines, with the largest gains on tasks involving multi-step quantitative criteria such as temporal filtering and threshold evaluation. The results also show that the framework's advantage diminishes on simpler diagnostic-lookup tasks, suggesting that operation-mechanism alignment is most valuable when clinical criteria demand heterogeneous reasoning operations that single-pass generation handles unreliably.

These findings support a general design principle: clinical decision support systems should match each reasoning operation to the mechanism whose failure modes are least harmful for that operation, rather than delegating all operations to a single generative model.

Limitations

Evaluation scope. Our evaluation covers six clinician-annotated binary decision tasks drawn from MIMIC-III, MIMIC-IV, and a de-identified partner-hospital dataset across four disease areas. Cohort sizes (50–151 patients per task) and the class imbalance in some tasks (e.g., 46/4 for asthma exacerbation risk) limit the precision of absolute per-task estimates; we therefore focus claims on relative comparisons across methods. Although clinicians were involved in task design and annotation, the auditability benefits of provenance traces remain to be evaluated systematically, including their effect on review efficiency, error detection, and decision review. Broader validation across additional institutions, languages, disease areas, and non-binary decision formats is a natural next step.

Manual task template construction. Task templates—including clinical criteria, node types, and dependency structure—are authored by domain experts for each task. This is a deliberate design choice: explicit expert-authored templates prioritize auditability, clinical face validity, and adherence to established guideline logic. At the same time, extending the framework to a new clinical task currently requires human authoring effort. Automating template construction from guideline text, clinical pathways, or existing decision-support artifacts is a promising extension that would improve scalability.

Mechanism-level ablation and error decomposition. Our experiments evaluate the framework holistically against end-to-end baselines, rather than isolating the contribution of each execution mechanism. A finer-grained ablation—holding the task graph fixed while varying which mechanism handles each node type, for instance routing computation nodes through an LLM rather than a deterministic tool—would more directly test the operation-mechanism alignment claim. We also do not report a per-node-type decomposition of residual errors, which would help localize failures to evaluation ambiguity, computation edge cases, or upstream error propagation. Both are natural follow-ups that the typed graph structure is de-

signed to support.

Empirical comparison with tool-using agents. We do not include a direct empirical comparison with autonomous tool-using agents on our task suite. As discussed above, we view such systems as complementary to operation-mechanism alignment rather than as substitutable baselines. We also note that "tool-using agents" is not a single configuration but a design space spanning tool selection, agent loop (e.g., ReAct, Toolformer, function-calling, code interpreter), planning strategy, error handling, and stopping criteria, each of which can substantially affect performance. Designing a comparison that controls these factors fairly is a non-trivial undertaking that we leave to future work.

Computational cost and latency. Because the framework decomposes a single decision into multiple typed nodes executed across heterogeneous mechanisms, end-to-end inference incurs higher computational cost and latency than single-pass prompting, which we do not quantify in this work. Systematic characterization of the accuracy–cost trade-off—including strategies for adaptive depth or selective decomposition—is an important practical follow-up for deployment settings where throughput and API cost are binding constraints.

Acknowledgments

We gratefully acknowledge Cone Health for providing the de-identified, real-world electronic health record (EHR) data used in this study.

References

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2025. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1):AIdbp2500418.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

William J Evans, John E Morley, Josep Argilés, Connie Bales, Vickie Baracos, Denis Guttridge, Aminah Jatoi, Kamyar Kalantar-Zadeh, Herbert Lochs, Giovanni Mantovani, and 1 others. 2008. Cachexia: a new definition. *Clinical nutrition*, 27(6):793–799.

Kenneth Fearon, Florian Strasser, Stefan D Anker, Ingvar Bosaeus, Eduardo Bruera, Robin L Fainsinger, Aminah Jatoi, Charles Loprinzi, Neil MacDonald, Giovanni Mantovani, and 1 others. 2011. Definition and classification of cancer cachexia: an international consensus. *The lancet oncology*, 12(5):489–495.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. 2023. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286.

Global Initiative for Asthma. 2024. Global strategy for asthma management and prevention. 2024 update. Accessed: 2026-04-17.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents, 2025. *URL https://arxiv. org/abs/2501.14654*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, volume 2025, pages 20715–20735.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. 2025. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi A Omiye, Mehr Kashyap, and 1 others. 2024. Factehr: A dataset for evaluating factuality in clinical notes using llms. *arXiv preprint arXiv:2412.12422*.

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, and Rex Ying. 2025. Experience retrieval-augmentation with electronic health records enables accurate discharge qa. *arXiv preprint arXiv:2503.17933*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

G Xiong, Q Jin, X Wang, M Zhang, Z Lu, and A Zhang. 2024a. Improving retrieval-augmented generation in medicine with iterative follow-up questions. arxiv; 2024. *arXiv preprint arXiv:2310.19988*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024b. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Parastoo Falakaflaki, Elena Kayayan, Guanyu Tao, Mahta Haghighat Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, and 1 others. 2024. The potential and pitfalls of using a large language model such as chatgpt, gpt-4, or llama as a clinical assistant. *Journal of the American Medical Informatics Association*, 31(9):1884–1891.

A Model Sensitivity Under Direct Prompting

Table 2 examines sensitivity to model choice under the direct-prompting setting. Unlike Table 1, which fixes GPT-5.4 and compares reasoning strategies under a controlled setup, this analysis varies the underlying model while keeping the prompting paradigm unchanged.

Performance under direct prompting varies across models and tasks. On COPD exacerbation evaluated on MIMIC-III, all three models have identical specificity (.349), suggesting that the false-positive pattern is a limitation of direct prompting rather than a model-specific weakness. On the same task evaluated on Cone Health, by contrast, the three models occupy different points on the sensitivity-specificity trade-off (GPT-5.2 conservative at .556/.951, GPT-5.4 balanced at .778/.780, Claude Opus 4.6 over-predicting at .722/.756); the framework in Table 1 (.833/.988) dominates all three on both dimensions, indicating that the gains are not attributable to a particular model's operating point. On the cachexia tasks, between-model differences are smaller than the improvement from replacing direct prompting with our structured framework. On the CKD tasks, direct-prompting performance is more consistent across models, but no single model reaches the precision or F1 of the structured framework.

Overall, the effect of model choice is task-dependent. We therefore interpret Table 2 as a sensitivity analysis showing that changing the underlying model does not by itself resolve the limitations of single-pass prompting, and that the gains in Table 1 cannot be attributed to model substitution alone.

Task	Model	Sens.	Spec.	Prec.	F1	Acc.
COPD Exacerbation (MIMIC-III)	GPT-5.2	.843	.349	.765	.802	.702
	GPT-5.4	.843	.349	.765	.802	.702
	Claude Opus 4.6	.870	.349	.771	.817	.722
COPD Exacerbation (Cone Health)	GPT-5.2	.556	.951	.714	.625	.870
	GPT-5.4	.778	.780	.438	.560	.780
	Claude Opus 4.6	.722	.756	.394	.510	.750
Asthma Risk[†]	GPT-5.2	.913	.250	.933	.923	.860
	GPT-5.4	.913	.000	.913	.913	.840
	Claude Opus 4.6	1.00	.000	.920	.958	.920
Cachexia (Fearon)	GPT-5.2	.800	.800	.632	.706	.800
	GPT-5.4	.800	.743	.571	.667	.760
	Claude Opus 4.6	.733	.743	.550	.629	.740
Cachexia (Evans)	GPT-5.2	.750	.595	.261	.387	.620
	GPT-5.4	.875	.452	.233	.368	.520
	Claude Opus 4.6	.625	.762	.333	.435	.740
Confirmed CKD	GPT-5.2	.720	.946	.818	.766	.880
	GPT-5.4	.760	.947	.826	.792	.900
	Claude Opus 4.6	.760	.960	.864	.809	.910
At-Risk CKD	GPT-5.2	.805	.862	.805	.805	.830
	GPT-5.4	.833	.862	.814	.824	.850
	Claude Opus 4.6	.857	.897	.857	.857	.880

Table 2: Model sensitivity under direct prompting. Performance varies across models and tasks, indicating that model choice matters, but this variation does not by itself resolve the limitations of single-pass prompting observed in Table 1. [†]High class imbalance (46/4).

MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP

Santiago Martínez Novoa[1], Lina María Gomez Meza[1], Juan Camilo Prieto[1], Ruben Manrique[1]
[1]Universidad de los Andes, Bogotá, Colombia
{s.martinezn, l.gomez1, jc.prietoa, rf.manrique}@uniandes.edu.co

Abstract

Despite Spanish being one of the most widely spoken languages in the world, biomedical NLP resources and systematic evaluations remain limited relative to English. We address this gap by constructing and releasing two Spanish biomedical corpora: (1) **MeSHClass-ES**, a 29,063 abstract bilingual corpus translated from PubMed with Opus-MT, and (2) **AnatEM-ES**, the AnatEM anatomical entity corpus translated with a chunk-level LLM-based pipeline that jointly preserves BIO annotations across 13,849 entity mentions. Both corpora achieve a mean COMET score of 0.73 despite using different translation systems. We benchmark nine encoder models spanning general-domain Spanish, domain-specific, and multilingual architectures for both tasks. RigoBERTa-2.0 leads both tasks (micro-F1 classification 0.69, tied with SciBETO-large; NER F1 0.66). Both domain pretraining and model capacity drive performance, with the gap slightly more pronounced for NER (4-point spread) than classification (3-point spread). XLM-RoBERTa-large emerges as a competitive multilingual baseline. A parallel evaluation of four open-weight decoders (7–9B) reveals a task-dependent encoder-decoder gap: QLoRA-adapted Gemma-2-9B reaches 88% of the best encoder on classification (micro-F1 .61 vs .69), but for NER every decoder configuration we tested stays at or below 40% of the best encoder F1. We release both corpora on the HuggingFace Hub[1], translation pipelines, and evaluation code on GitHub.[2]

1 Introduction

Biomedical NLP has advanced significantly through transformer-based models trained in large English corpora, allowing effective entity recognition (Lee et al., 2019), clinical prediction (Huang et al., 2020), and concept normalization (Wei et al., 2015). However, these advances remain largely inaccessible to Spanish-speaking communities: while English has 11.35 models per million speakers on HuggingFace, Spanish has only 2.30 (García Subies et al., 2024). This disparity is particularly acute in the biomedical domain, where dedicated models such as BioBERT (Lee et al., 2019), PubMedBERT (Gu et al., 2021) and BioGPT (Luo et al., 2022) exist for English, but Spanish equivalents are limited to a handful of initiatives (Carrino et al., 2021b,a).

Two questions remain open for Spanish biomedical NLP. First, which of the available Spanish and multilingual encoder architectures is best suited for biomedical tasks, and how much does domain-specific pretraining matter? Second, can open-weight instruction-tuned LLMs offer a viable alternative without task-specific fine-tuning?

We answer these questions on two complementary tasks: broad-topic MeSH classification (no open Spanish corpus exists at PubMed scale) and fine-grained anatomical NER. Existing native Spanish biomedical NER corpora (PharmaCoNER, DisTEMIST, SympTEMIST) target drugs, diseases, and symptoms; none provide the 12-type anatomical granularity of AnatEM (Pyysalo and Ananiadou, 2014), which makes translation the only route to an anatomy-focused Spanish NER benchmark at this granularity.

We address these questions with three contributions:

1. **Two Spanish biomedical corpora** built with translation approaches suited to each downstream task: MeSHClass-ES (44,604 abstracts translated with Opus-MT; 29,063 with MeSH labels of major-topics used for classification) and AnatEM-ES (1,212 documents, LLM-

[1]MeSHClass-ES: https://huggingface.co/datasets/Flaglab/pubmed_mesh_spanish AnatEM-ES: https://huggingface.co/datasets/Flaglab/anat-em-es

[2]https://github.com/SantiagoM99/spanish-medical-nlp

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 617–629
July 3-4, 2026 ©2026 Association for Computational Linguistics

based fragment translation with joint preservation of BIO-label and human verification) for NER. Both are COMET-assessed and are released in bilingual format.

2. **A systematic encoder benchmark** of nine models that span general-domain Spanish, domain-specific, and multilingual architectures for both tasks.

3. **Evaluation of a decoder** of four open-weight LLMs (7–9B parameters) in zero-shot, few-shot, retrieval-augmented, and QLoRA settings, demonstrating that current decoders cannot substitute for fine-tuned encoders.

Our work aligns with the growing calls for the evaluation of non-English biomedical NLP (Neves et al., 2023; García Subies et al., 2024) and contributes reproducible resources to the community.

2 Related Work

Biomedical NLP in Spanish. The development of Spanish biomedical NLP has been driven by two lines of work. On the resource side, the Barcelona Supercomputing Center released RoBERTa-BioMed-ES, pretrained on a 1B-token biomedical and clinical corpus (Carrino et al., 2021b). The underlying Spanish Biomedical Crawled Corpus (Carrino et al., 2021a) remains one of the few large-scale open resources for the domain. On the evaluation side, shared tasks such as PharmaCoNER (Gonzalez-Agirre et al., 2019), MEDDOCAN (Marimon et al., 2019), and DisTEMIST/SympTEMIST (Miranda-Escalada et al., 2022) have produced annotated corpora for clinical NER. A recent survey by García Subies et al. (2024) benchmarked 17 corpora across Spanish clinical encoder models, confirming the value of domain-specific pretraining, but noting the persistent gap between clinical and biomedical evaluation resources. RigoBERTa-2.0 (Vaca Serrano et al., 2022) adapted the XLM-RoBERTa-large parameters (∼560M) to Spanish through continued pretraining in a curated Spanish corpus and was subsequently extended to the clinical domain as RigoBERTa-Clinical (Subies et al., 2025), trained in the ClinText-SP corpus of Spanish clinical text. SciBETO (Flaglab, 2024) is the first encoder pretrained in Spanish scientific text, compiling 1.9M documents (18.2B tokens) from Colombian and international open-access repositories and training RoBERTa-base (125M) and large (355M) variants.

Benchmarking biomedical models. In English, comprehensive benchmarks such as BLURB (Gu et al., 2021) and recent LLM evaluations (Chen et al., 2025) have established baselines for NER, relation extraction, classification, and answer questions. Gade et al. (2025) evaluated 12 model architectures in 61 biomedical corpora for zero-shot relation triplet extraction, finding that even the best LLMs achieve F1 below 0.5, while fine-tuned encoders still dominate in their training corpora. For non-English settings, BioMistral (Labrak et al., 2024) included multilingual evaluation but relied on auto-translated English benchmarks, leaving open questions about performance on domain-translated tasks.

MeSH classification. Automatic MeSH indexing has been driven by the BioASQ challenge series (Tsatsaronis et al., 2015), with representative approaches including DeepMeSH, which couples learned sparse and dense representations for large-scale multi-label indexing (Peng et al., 2016), and self-attentive architectures such as MeSHProbeNet (Xun et al., 2019). All of these efforts are targeted at English PubMed. No comparable Spanish evaluation setup has been released, which MeSHClass-ES addresses.

Translation-based corpus construction. Neural machine translation has been used to create biomedical resources in low-resource languages, with Helsinki-NLP Opus-MT models (Tiedemann and Thottingal, 2020) serving as a common backbone. For span-annotated corpora, label preservation throughout translation is typically handled by annotation projection (Yarowsky et al., 2001), which decouples translation from alignment and can fail in domain text where aligner coverage is poor. Recent work has systematically compared these approaches in clinical settings: Gaschi et al. (2023) benchmarked cross-lingual transfer against translate-train and translate-test strategies for clinical NER in French and German, showing that translation-based methods can match cross-lingual transfer but require careful alignment design; Lanz and Pecina (2025) applied a similar translation pipeline with answer projection for clinical question answering across six languages, finding that multilingual models can compete with monolingual domain-specific ones. We instead use an LLM-based chunk translation that preserves BIO tags via instruction, trading explicit alignment for the model's joint handling of translation

and tagging (§3.2). Previous work has validated translated biomedical corpora through BLEU (Papineni et al., 2002) and more recently COMET (Rei et al., 2020), a learned metric that correlates more strongly with human judgments for domain text; we adopt COMET-based validation following these recommendations.

3 Dataset Construction

We construct and release two Spanish biomedical corpora, each built with a translation approach suited to its downstream task. MeSHClass-ES consists of plain abstracts with document-level MeSH labels: sentence-level neural MT with Opus-MT is sufficient, cheap at 44,604 document scale, and leaves labels untouched. AnatEM-ES is token-annotated with BIO tags, which makes naive sentence MT insufficient: we use a chunk-level LLM pipeline (GPT-4o) that jointly translates the text and preserves the BIO labels via instruction, avoiding explicit cross-lingual alignment at the cost of a more expensive per-document call. Figure 1 illustrates both pipelines. Both corpora are distributed in bilingual format, retaining the original English text alongside the Spanish translations, enabling future re-translation with improved MT systems without repeating the source collection and curation pipeline. Both corpora will be released on the HuggingFace Hub.

3.1 MeSHClass-ES corpus

Source selection. We retrieve open-access abstracts in Spanish from PubMed via the NCBI E-utilities API, combining three complementary search strategies to overcome the 9,999-result API limit: year-by-year queries (1988–2025), queries over 20 Spanish-language biomedical journals, and queries over 35 MeSH subject headings. All queries require `spanish[Language]`, `hasabstract`, and open-access filters. From 78,733 initial hits, we retrieve 49,039 unique PMIDs, filtered to yield a final corpus of **44,604 articles** with complete metadata (title, abstract, MeSH descriptors, journal, publication year). Of these, **29,063** carry at least one major-topic MeSH descriptor (`MajorTopicYN="Y"`) and form the classification benchmark; the remaining articles are retained in the bilingual release but excluded from classification, as non-major descriptors are too broad to serve as supervisory labels. The top journals and the full collection funnel are reported in Appendix C.

Translation. We translate abstracts and titles into Spanish using `Helsinki-NLP/opus-mt-en-es` (Tiedemann and Thottingal, 2020) sentence by sentence. We chose Opus-MT for its open-source availability, reproducibility, and established use in prior translation-based corpus construction (Gaschi et al., 2023). We did not compare alternative MT systems (e.g., NLLB, MadLad) for MeSHClass-ES; however, the convergence of COMET scores between Opus-MT and GPT-4o on AnatEM-ES (both 0.73) despite their vastly different architectures and training data suggests that absolute translation quality is bounded by domain difficulty rather than MT system choice (Section 6). The bilingual release format further mitigates this limitation by enabling re-translation with improved systems without repeating the collection pipeline. All 44,604 articles were successfully translated (166 were already in Spanish and bypassed translation). The mean abstract length expands from 193.5 words in English to 224.2 words in Spanish ($\sim$1.16$\times$ expansion), consistent with typical EN$\rightarrow$ES translation ratios.

Translation quality. We evaluated a random sample of 200 abstracts with COMET (`Unbabel / wmt22-comet-da`) (Rei et al., 2020), a neural metric trained on human quality judgments. Our Opus-MT translations achieve a mean score of 0.73 (SD = 0.08; median = 0.74; range: 0.22–0.85), with 38% above 0.75 and 6% above 0.80. These values indicate moderate quality; following Freitag et al. (2022), we use COMET rather than BLEU as it provides more reliable quality estimates for domain text. The low proportion above 0.80 reflects the inherent difficulty of biomedical translation, where specialized terminology, nominal compounds, and abbreviations challenge general-domain MT. We discuss downstream implications in Section 6.

Corpus statistics. Figures 2–3 show the temporal distributions of the MeSH category. Health Care (N) and Techniques (E) dominate; Anatomy (A), Humanities (K), and Technology/Industry (J) are the rarest, producing a 9:1 imbalance ratio.

3.2 AnatEM-ES corpus

Translation with joint annotation preservation. The AnatEM corpus (Pyysalo and Ananiadou, 2014) (1,212 documents, 13,701 anatomical entity mentions across 12 types) is distributed in token-per-line BIO format, which makes naive transla-

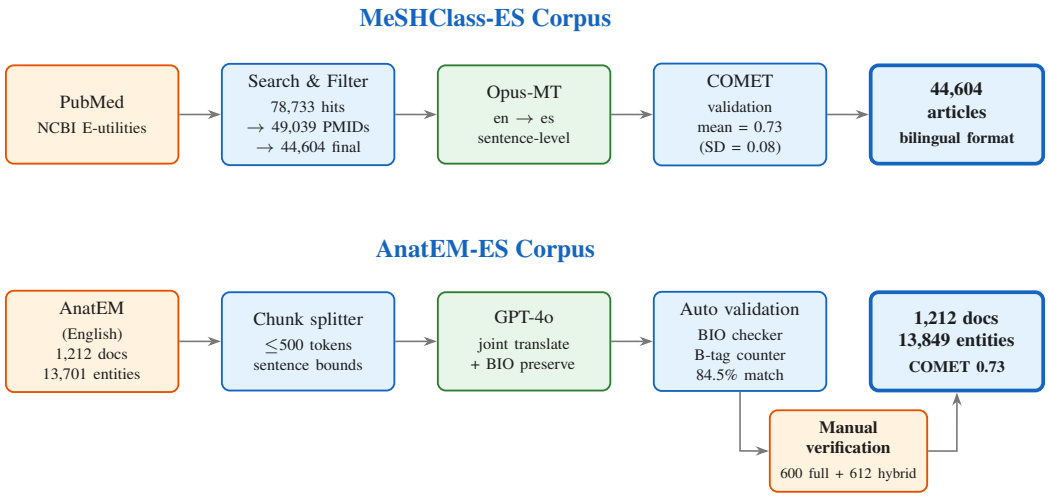


Figure 1: Construction pipelines for MeSHClass-ES (top) and AnatEM-ES (bottom). The two corpora use different translation approaches matched to their annotation requirements: sentence-level neural MT for plain-text abstracts, and chunk-level LLM-based translation with joint BIO-label preservation for token-annotated documents.

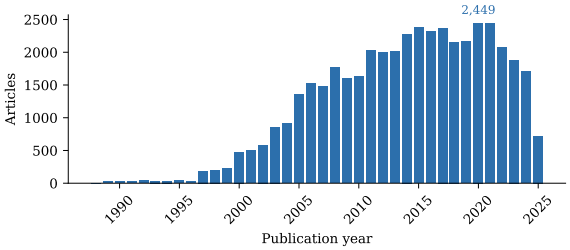


Figure 2: Publication year distribution of MeSHClass-ES articles (1988–2025). Peaks around 2015 and 2020–2021 reflect growth in open-access Spanish biomedical publishing.

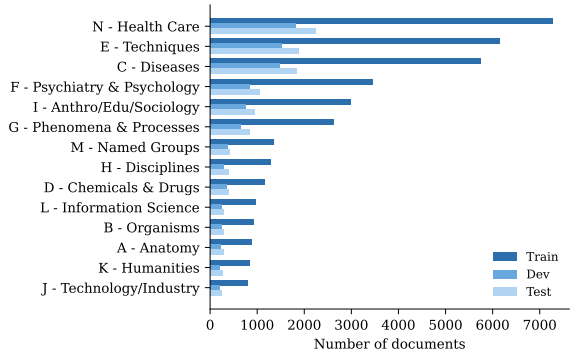


Figure 3: MeSH Level-1 category distribution across train/dev/test splits. Categories V (Publication Type) and Z (Geography) are excluded.

tion problematic: splitting by token loses linguistic context, while translating plain text loses the alignment between tokens and labels. Standard annotation projection via word aligners (Dou and Neubig, 2021) assumes roughly monotonic token correspondences, which breaks down when EN→ES translation introduces word reordering or inserts prepositions within entity spans (e.g., "blood samples" → "muestras de sangre", where the inserted preposition "de" must receive an I-tag despite having no source-side counterpart). We therefore adopt a chunk-level LLM-based pipeline using GPT-4o via the OpenAI Batch API that sidesteps explicit alignment by handling translation and re-labeling as a single operation. Each document is split into chunks of up to 500 tokens at natural sentence boundaries (preserving document integrity), and the model is instructed via a domain-specific prompt (Appendix A) to (i) translate the text into fluent Spanish, (ii) keep the original BIO labels attached to the translated tokens, and (iii) preserve the 12 entity types. This trades aligner-based guarantees for instruction-following fidelity, at the cost of relying on the LLM's ability to jointly manage translation and tagging. We do not compare the two strategies head-to-head; such a comparison would be a valuable direction for future work.

Validation. Two automatic validators check all outputs: a BIO consistency checker (orphan I-tags, type mismatches, invalid formats) and a B-tag count comparator against the English original. The translated corpus contains **13,849 entity mentions** (+1.1% vs. English), with exact counts matching in 1,024/1,212 documents (84.5%); typical discrepancies are occasional duplicate segments or minor label-formatting variants. One author manually verified the first 600 documents for BIO conven-

Entity Type	Train	Dev	Test
Anatomical_system	67	7	39
Cancer	1,469	513	1,129
Cell	2,459	666	1,447
Cellular_component	413	138	287
Developing_anat_str.	45	20	35
Immaterial_anat_ent.	116	50	97
Multi-tissue_str.	884	274	559
Organ	413	146	306
Organism_subdiv.	166	62	109
Organism_subst.	338	102	244
Pathological_form.	182	63	151
Tissue	464	129	260
Total	7,016	2,170	4,663

Table 1: AnatEM-ES entity type distribution (606/202/404 documents in train/dev/test).

tions, entity typing, and fluency; the remaining 612 followed a hybrid protocol prioritizing documents with the largest number of flagged inconsistencies, applying LLM-assisted correction at scale, and a second manual pass on critical cases.

Translation quality. All 1,212 documents are scored with COMET (`Unbabel/wmt22-comet-da`); AnatEM-ES reaches mean 0.73 (SD 0.06; range 0.35–0.99; 36% above 0.75, 8% above 0.80). These scores are nearly identical to MeSHClass-ES (0.73) despite using a different MT system (GPT-4o vs. Opus-MT), suggesting a domain-level quality ceiling rather than a system-specific limitation.

Entity distribution. Table 1 shows the entity type distribution in AnatEM-ES. Cell mentions dominate (35%), followed by Cancer (22%) and Multi-tissue_structure (13%). Rare types such as Developing_anatomical_structure (45 train mentions) and Anatomical_system (67 train) present challenges for fine-grained NER evaluation.

3.3 Corpus quality characterization

We characterize both corpora as silver-standard resources with respect to translation quality. For MeSHClass-ES, the supervisory labels are gold-standard (assigned by trained NLM indexers to the original English articles) and are not affected by translation, but the Spanish text is machine-translated without systematic human post-editing. Level-1 MeSH categories are derived by taking the first letter of each MeSH descriptor's tree number (e.g., C01.539 maps to category C, Diseases). For AnatEM-ES, entity annotations have been verified through a combination of full manual review (600 documents) and a hybrid protocol targeting flagged inconsistencies (612 documents); while this exceeds typical silver-standard corpora, residual annotation noise from the translation process likely persists in a subset of documents. We encourage users to consider the bilingual format for cross-referencing against the original English annotations when high-fidelity labels are required.

3.4 Task formulations

MeSH classification. Multi-label classification over 14 MeSH Level-1 categories (A through N, excluding Geography and Publication Type). We restrict the task to the 29,063 articles with at least one MajorTopic MeSH descriptor (`MajorTopicYN="Y"`), since non-major descriptors are too broad to serve as supervisory labels, and apply a random split (18,600 / 4,650 / 5,813 train/dev/test, approximately 64/16/20; mean 1.96 labels per document). All models receive title + abstract concatenation.

AnatEM NER. Token-level BIO tagging over 12 anatomical entity types, using the original AnatEM splits: 606 / 202 / 404 train/dev/test documents.

4 Experimental Setup

4.1 Models

Encoder fine-tuning. We evaluated nine encoders in three categories: *general-domain Spanish* (BETO (Cañete et al., 2023), BERTIN (de la Rosa et al., 2022), and RigoBERTa-2.0 (Vaca Serrano et al., 2022)); *specific-domain* (SciBETO-base/large (Flaglab, 2024), RigoBERTa-Clinical (Subies et al., 2025), RoBERTa-Bio-Clinical-ES (Carrino et al., 2021b)); and *multilingual* (XLM-RoBERTa base / large (Conneau et al., 2020)). We use `AutoModelForSequenceClassification` for classification and `AutoModelForTokenClassification` with BIO labels and first-piece sub-word alignment for NER.

Decoder zero/few-shot. Four open-weight instruction-tuned LLMs: Gemma-2-9B-it (Team et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), all loaded in 4-bit NF4 quantization. For classification, we use zero-shot and few-shot (3 examples) prompts in Spanish. For NER, we additionally test the k-NN few-shot retrieval

Hyperparameter	Search space
Learning rate	{1e-5, 2e-5, 3e-5}
Batch size	{16, 32}
Max epochs	10
Warmup ratio	0.06
Weight decay	0.1

Table 2: Encoder fine-tuning hyperparameter search space, applied to both MeSHClass-ES and AnatEM-ES.

(k = 5) using FAISS with `all-MiniLM-L6-v2` embeddings (Reimers and Gurevych, 2019), and optional self-verification (Wang et al., 2023) in which the model re-examines its own predictions. All prompts are provided in the Appendix A.

PEFT. We fine-tuned all four decoders with QLoRA (Dettmers et al., 2023) (r = 16, α = 32, dropout=0.05) in attention projection layers, NF4 quantization, AdamW 8-bit. The NER QLoRA prompt emits a linear space-separated BIO tag sequence (one tag per input token), matched between training and inference; this differs from the JSON format used by zero-shot and few-shot prompts (see Appendix A).

4.2 Training configuration

For encoder fine-tuning, we grid-search over the hyperparameters in Table 2 (the same space for both tasks), with maximum sequence length 512, AdamW FP16, sigmoid + threshold 0.5 for classification. Checkpoints are selected by validation loss and test F1 is reported once per model from the best-validation-loss configuration, so the test split is not used for model selection. QLoRA runs use 3 epochs, effective batch size 16 (4× gradient accumulation), lr = 2e-4. Decoder inference uses greedy decoding with `max_new_tokens` 100 (cls) / 512 (NER).

4.3 Evaluation metrics

Classification: F1-score (micro and macro); micro precision and recall are also reported for decoders. NER: entity-level precision, recall, and F1-score via `seqeval` (Nakayama, 2018) with strict span matching.

5 Results

5.1 Multi-label classification

Table 3 presents encoder classification results. All models fall within a narrow 3-point micro-F1 range (0.67–0.69), suggesting that the task is moderately tractable for fine-tuned encoders. RigoBERTa-2.0 leads both metrics (micro-F1 0.692, macro-F1 0.584), tied on micro-F1 by SciBETO-large (0.692, macro-F1 0.577); XLM-RoBERTa-large follows at micro-F1 0.688. Among the domain-specific models, RigoBERTa-Clinical is the closest to the leaders on macro-F1 (0.580), consistent with its clinical pretraining giving it broader coverage of tail categories. BERTIN and BETO occupy ranks 5 and 7 of 9, suggesting that classification relies more on topic-level semantics than on domain vocabulary.

Model	Mi-F1	Ma-F1
RigoBERTa-2.0	**.692**	**.584**
SciBETO-large	.692	.577
XLM-R (large)	.688	.578
RigoBERTa-Clinical	.681	.580
BERTIN	.679	.573
SciBETO-base	.678	.561
BETO	.675	.553
RoBERTa-Bio-Clin.	.675	.557
XLM-R (base)	.666	.556

Table 3: Encoder multi-label MeSH classification results (test set, T+A input). Best per column in **bold**; RigoBERTa-2.0 and SciBETO-large tie at the reported precision on micro-F1 (0.6921 vs. 0.6917 at four decimals).

Table 4 presents decoder results. The zero-shot best decoder (Gemma-2-9B, micro-F1 .458) reaches 66% of the best encoder. Few-shot prompting hurts micro-F1 for three of four decoders (Gemma −3 pts, Llama −4, Qwen −8), with Mistral as the exception (+3 pts from .355 to .384); macro-F1, in contrast, improves mildly across all four (+1 to +5 pts), indicating that fixed exemplars help tail-category coverage even as they hurt overall precision by anchoring models toward the exemplar label combinations. QLoRA adaptation tells a very different story: Gemma-2-9B reaches micro-F1 .608 / macro-F1 .500, within 8 points of the best encoder (micro-F1 .692) and corresponding to ∼88% of encoder performance; Mistral-7B follows at .535 / .450. Llama (.427) and Qwen (.416) trail on micro-F1, but all four models show sharply improved macro-F1 (+14–26 points over zero-shot), indicating that PEFT helps with tail-category coverage across the board. Model choice matters substantially under QLoRA: the 9B Gemma and 7B Mistral outperform Llama and Qwen by 11–19 micro-F1 points with the same training recipe, narrowing the encoder–decoder gap considerably for classification.

Setting	Model	Mi-P	Mi-R	Mi-F1	Ma-F1
Zero-shot					
	Gemma-2-9B	.352	**.654**	**.458**	**.282**
	Llama-3.1-8B	.398	.459	.426	.178
	Qwen2.5-7B	**.401**	.438	.418	.216
	Mistral-7B	.259	.564	.355	.190
Few-shot (3 ex.)					
	Gemma-2-9B	**.325**	**.615**	**.426**	**.289**
	Llama-3.1-8B	.317	.497	.387	.217
	Qwen2.5-7B	.292	.412	.342	.234
	Mistral-7B	.284	.591	.384	.244
QLoRA					
	Gemma-2-9B	**.619**	**.597**	**.608**	**.500**
	Llama-3.1-8B	.564	.343	.427	.367
	Qwen2.5-7B	.593	.320	.416	.358
	Mistral-7B	.611	.475	.535	.450

Table 4: Decoder classification results (T+A input, greedy decoding). Best per column in **bold**.

Model	P	R	F1
RigoBERTa-2.0	**.661**	.661	**.661**
RigoBERTa-Clinical	.638	**.664**	.651
XLM-R (large)	.640	.655	.647
SciBETO-large	.638	.648	.643
RoBERTa-Bio-Clin.	.644	.622	.633
BETO	.624	.630	.627
BERTIN	.620	.622	.621
XLM-R (base)	.616	.621	.618
SciBETO-base	.612	.623	.618

Table 5: Encoder NER results (entity-level, strict span match) on the translated AnatEM-ES test set. Best per column in **bold**.

5.2 Named entity recognition

Table 5 shows the NER encoder results. The spread is 4 F1 points, slightly larger than for classification. RigoBERTa-2.0 achieves the best F1 (0.661), ahead of RigoBERTa-Clinical (0.651) and XLM-RoBERTa-large (0.647). SciBETO-large (0.643) and RoBERTa-Bio-Clinical-ES (0.633) form the next tier, both benefiting from domain-focused pretraining. BETO (0.627), BERTIN (0.621), XLM-RoBERTa-base and SciBETO-base (both 0.618) trail behind. While RigoBERTa-2.0 leads both tasks, the mid-table reshuffles substantially: SciBETO-large drops from a tie for 1st on classification to 4th on NER, and RigoBERTa-Clinical rises from 4th to 2nd, reinforcing that task-specific evaluation is essential.

Table 6 presents decoder NER results. The encoder-decoder gap is even more pronounced than in classification. Gemma-2-9B achieves the best decoder F1 (0.27), roughly 1.8× better than other decoders but still only 40% of the best encoder. Contrary to expectations, k-NN retrieval-augmented prompting substantially *degrades* performance: across all four decoders, recall collapses from .13–.30 to $\leq$.012 while precision stays moderate (.17–.27), suggesting that retrieved examples encourage pattern copying rather than entity boundary detection, a failure mode amplified by translation-induced span inconsistencies. Adding a self-verification step raises precision sharply (.23–.57) but does not recover recall, so F1 stays at $\leq$.023 for every decoder. QLoRA-adapted decoders collapse entirely (F1 $\leq$ .004 for all four models), indicating that the generative BIO-labeling formulation is fundamentally unsuited for 7–9B parameter decoders under these training conditions; the gap between the best and worst QLoRA decoder is within rounding error, so model choice does not rescue the formulation here.

Setting	Model	P	R	F1
Zero-shot				
	Gemma-2-9B	**.243**	**.295**	**.267**
	Llama-3.1-8B	.117	.211	.151
	Qwen2.5-7B	.132	.126	.129
	Mistral-7B	.115	.129	.122
k-NN few-shot (k=5)				
	Gemma-2-9B	.228	**.012**	**.024**
	Llama-3.1-8B	**.266**	**.012**	.023
	Qwen2.5-7B	.211	.007	.013
	Mistral-7B	.171	.004	.009
k-NN + self-verification				
	Gemma-2-9B	.471	**.012**	**.023**
	Llama-3.1-8B	**.567**	.004	.007
	Qwen2.5-7B	.357	.006	.013
	Mistral-7B	.230	.004	.008
QLoRA				
	Gemma-2-9B	.002	**.006**	.003
	Llama-3.1-8B	.000	.000	.000
	Qwen2.5-7B	.000	.001	.001
	Mistral-7B	**.005**	.004	**.004**

Table 6: Decoder NER results (entity-level, strict span match). Best F1 per column in **bold**. k-NN uses FAISS with `MiniLM-L6-v2` embeddings over the training set.

6 Analysis and Discussion

Domain pretraining helps, but the advantage is task-dependent. For classification, the spread among encoders is 3 points (micro-F1 .67–.69), with general-domain base models like BETO and BERTIN performing near the domain-specialized ones. For NER, the spread is 4 points (.62–.66), with the top group (RigoBERTa-2.0, RigoBERTa-Clinical, XLM-RoBERTa-large, SciBETO-large) clearly separating from base-sized models. Classi-

fication relies more on topic-level semantics available in general pretraining; NER benefits from both model capacity and domain-specific entity coverage. Practically, choosing RigoBERTa-2.0 over BERTIN yields a 4-point F1 gain on NER but only about 1 point on classification, meaning model selection matters more for token-level tasks. We note a confound: in our model set capacity and domain pretraining are partially correlated — all four NER top performers are large models, and the only base-sized domain model (RoBERTa-Bio-Clinical-ES) ranks 5th. We cannot cleanly separate the two factors with the current pool, and a stronger claim would require a large biomedical-pretrained Spanish encoder of comparable size, which is not yet available.

XLM-RoBERTa-large is a strong multilingual baseline. Despite lacking Spanish-specific or scientific pretraining, XLM-RoBERTa-large ranks 3rd on classification (micro-F1 .688, within 0.5 points of the leaders) and 3rd on NER (F1 .647, 1.4 points behind RigoBERTa-2.0). Its competitive macro-F1 (.578) in classification indicates that cross-lingual pretraining on 100 languages provides robust coverage of tail categories. This makes it a practical default when domain-specific Spanish models are unavailable.

Encoder-decoder gap depends on task and adaptation. Zero-shot, Gemma-2-9B reaches 66% of the best encoder on classification and 40% on NER. The asymmetry reflects task nature: LLMs handle coarse topic classification reasonably but struggle with precise span identification. PEFT changes the picture sharply on classification (QLoRA Gemma 88%, Mistral 77%) but not on NER, where every decoder configuration stays below F1 .27.

Few-shot prompting degrades micro-F1 but helps macro-F1. Few-shot examples hurt micro-F1 for three of four decoders (Gemma -3, Llama -4, Qwen -8), with Mistral the exception ($+3$). The fixed exemplars (label combinations {C,D}, {C,N}, {A,C,E,M}) drive precision drops larger than recall drops in the three degrading cases, consistent with over-prediction of exemplar categories. Macro-F1, in contrast, improves across all four models ($+1$ to $+5$), suggesting that even biased exemplars surface tail categories that would otherwise be ignored. Static few-shot examples can thus be counterproductive for micro-F1 in multi-label settings unless the exemplars span the full label space.

QLoRA works for classification but fails for generative NER. For classification, QLoRA improves macro-F1 by $+14$–26 points over zero-shot across all four decoders, and Gemma/Mistral close most of the encoder gap in micro-F1. For NER, all four collapse to F1 $\leq .004$ despite training and evaluating in the same BIO format (§A): models emit all-O predictions (Llama, Qwen) or malformed tag strings (Gemma, Mistral). The generative BIO formulation requires one tag per token with valid B/I dependencies, which at 7–9B and 3-epoch, r=16 QLoRA none of the four decoders learn; longer training, higher-rank adapters, or a span-extraction output (e.g., JSON) would likely be needed. This aligns with English findings (Gade et al., 2025) but shows a wider gap, likely amplified by limited Spanish instruction-tuning data.

Translation quality likely caps absolute performance. COMET scores (mean 0.73 for both corpora, consistent across Opus-MT and GPT-4o) suggest a domain-level quality ceiling that bounds absolute F1, especially on NER where entity boundaries must align across languages. Since all models process the same translated text, relative rankings should be less affected. A cross-benchmark comparison on a native Spanish corpus would be the natural next step to verify this; however, no existing Spanish biomedical NER corpus covers anatomical entities at the granularity of AnatEM-ES. The closest available resource, PharmaCoNER, targets pharmacological entities and clinical text, making it unsuitable as a direct comparison point. This gap is precisely the motivation for the present work.

Decoder NER error regimes. Gemma-2-9B leads the decoder pool on NER (F1 .27, $\sim$1.8$\times$ the next best), likely reflecting its larger effective capacity, instruction-tuning data, or biomedical coverage. Manual inspection on a 449-sentence sample shows that the encoder–decoder gap is qualitative, not just quantitative. Zero-shot decoders over-predict: Llama emits 8,393 predictions for 4,669 gold mentions, and Mistral hallucinates entities in 51 sentences with no anatomical content (e.g., labeling "Cuba"/"Florida" as `Organism_subdivision`). Even on correct spans, all four models systematically mistype them (e.g., "arteria coronaria derecha" as `Organ` instead of `Multi-tissue_structure`). The k-NN and QLoRA settings exhibit the opposite regime: 89–

97% of sentences receive zero predictions and recall drops to 0.003–0.012, consistent with pattern-copying of retrieved exemplars and all-O collapse under QLoRA.

7 Conclusion

We released MeSHClass-ES (44,604 abstracts; 29,063 with major-topic MeSH labels) and AnatEM-ES (1,212 documents, 13,849 entities), two Spanish biomedical corpora built with translation pipelines matched to each task. RigoBERTa-2.0 leads both benchmarks, domain pretraining and capacity drive performance (more so on NER), and XLM-RoBERTa-large is a strong multilingual baseline. The encoder–decoder gap is task-dependent: QLoRA-adapted Gemma-2-9B reaches 88% of the best encoder on classification, but NER remains out of reach for all four 7–9B decoders. The bilingual corpus format enables re-translation as MT improves, without repeating collection.

Limitations

Our corpora are constructed via machine translation rather than from native Spanish biomedical text. COMET evaluation indicates moderate translation quality for both corpora (mean 0.73 for both MeSHClass-ES and AnatEM-ES, with 6–8% of documents above the 0.80 threshold), reflecting the difficulty of translating biomedical terminology regardless of the MT system used; this quality ceiling likely caps absolute downstream performance. For AnatEM-ES, the LLM-based translation with joint BIO preservation avoids explicit cross-lingual alignment but relies on the LLM's instruction-following fidelity: 15.5% of documents show entity-count discrepancies, and while manual verification covered the first 600 documents in full plus the most critical cases in the remaining 612, residual annotation noise likely persists. We evaluated two translation systems (Opus-MT and GPT-4o) but did not compare them head-to-head. Our encoder benchmark covers a grid over three learning rates and two batch sizes; Bayesian optimization could improve individual results. The decoder evaluation is limited to 7–9B parameter models; larger models may narrow the gap.

Ethical Considerations

All data are derived from publicly available PubMed abstracts and the open AnatEM corpus. Patient data or personally identifiable information are not involved. Our translation pipelines use Opus-MT (open-source) and the OpenAI Batch API (GPT-4o); the output is released together with the translation scripts. Machine-translated medical text should not be used for clinical decision-making without expert review.

References

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021a. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *Preprint*, arXiv:2109.07765.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021b. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *Preprint*, arXiv:2109.03570.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *Preprint*, arXiv:2308.02976.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Preprint*, arXiv:2207.06814.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. pages 2112–2128.

Flaglab. 2024. Scibeto collection. https://huggingface.co/collections/Flaglab/scibeto. Accessed: 2026-04-17.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Frederik Gade, Ole Lund, and Marie Lisandra Mendoza. 2025. Benchmarking zero-shot biomedical relation triplet extraction across language model architectures. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 88–100, Viena, Austria. Association for Computational Linguistics.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2024. A comparative analysis of spanish clinical encoder-based models on NER and classification tasks. *J. Am. Med. Inform. Assoc.*, 31(9):2137–2146.

Félix Gaschi, Xavier Fontaine, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical NER: Translation or cross-lingual transfer? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 289–311, Toronto, Canada. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Antonio Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, da Lisa Jaramillo Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Conference and Labs of the Evaluation Forum*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12):i70–i79.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. Clintext-sp and rigoberta clinical: a new set of open resources for spanish clinical nlp. *Preprint*, arXiv:2503.18594.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Alejandro Vaca Serrano, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto, and Alvaro Barbero Jimenez. 2022. Rigoberta: A state-of-the-art language model for spanish. *Preprint*, arXiv:2205.10233.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015:918710.

Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. MeSHProbeNet: a self-attentive probe net for MeSH indexing. In *Bioinformatics*, volume 35, pages 3794–3802.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*. Association for Computational Linguistics.

A Prompts

Classification: Zero-shot

```
Eres un experto en indexación biomédica
con conocimiento profundo de MeSH.

Categorías MeSH disponibles:
  A: Anatomía  B: Organismos
  C: Enfermedades
  D: Sustancias Químicas y Fármacos
  E: Técnicas Analíticas, Diagnósticas
     y Terapéuticas
  F: Psiquiatría y Psicología
  G: Fenómenos y Procesos
  H: Disciplinas y Ocupaciones
  I: Antropología, Educación y Sociología
  J: Tecnología, Industria y Agricultura
  K: Humanidades
  L: Ciencias de la Información
  M: Grupos Nombrados
  N: Cuidados de Salud
  V: Características de Publicación

Clasifica el siguiente abstract médico
asignándole los códigos MeSH de nivel 1.
Responde ÚNICAMENTE con los códigos
separados por comas (ejemplo: A, C, D).

Abstract: {text}
Códigos MeSH:
```

NER: Zero-shot

```
Eres un sistema experto en reconocimiento
de entidades de anatomía médica en español.

Tipos válidos: {entity_types}

Extrae TODAS las entidades del texto.
Responde SOLO con un array JSON:
[{"entity": "texto", "type": "tipo"}]
Si no hay entidades, responde: []

Texto: "{sentence}"
JSON:
```

NER: k-NN few-shot

```
[Same header as zero-shot]

EJEMPLOS (ordenados por similitud):
{retrieved examples from FAISS index}

Texto: "{sentence}"
JSON:
```

AnatEM-ES: Translation prompt (GPT-4o)

```
Eres un experto etiquetador médico
especializado en anatomía. Tu tarea es
traducir este dataset de entidades
anatómicas del inglés al español
manteniendo las etiquetas originales
en inglés.

REGLAS FUNDAMENTALES DE NER:
- B-[etiqueta]: Marca el INICIO de una
  entidad anatómica
- I-[etiqueta]: Marca la CONTINUACIÓN
  de la misma entidad
- O: Marca tokens que NO son entidades
  anatómicas
- NUNCA puede haber I- sin un B- previo
  de la misma etiqueta

INSTRUCCIONES CRÍTICAS:
1. Traduce SOLO las palabras, MANTÉN
   etiquetas en inglés
2. El número de etiquetas B- debe ser
   EXACTAMENTE igual al original
3. Si inviertes orden de palabras,
   ajusta etiquetas coherentemente
4. Si añades preposiciones/artículos
   para naturalidad:
   - Dentro de entidad: usar I- de la
     misma etiqueta
   - Fuera de entidad: usar O
5. No añadas etiquetas a términos que
   no tenían previamente
6. Es posible que haya textos donde hay
   ninguna etiqueta, no es necesario
   crear nuevas etiquetas.

EJEMPLOS CLAVE:

1. Inversión con preposición añadida:
ENTRADA:
cell       B-Organism_substance
lysate     I-Organism_substance
experiment O

SALIDA:
lisado     B-Organism_substance
celular    I-Organism_substance
experimento O

2. Adición de preposición dentro de
   entidad:
ENTRADA:
blood   B-Organism_substance
samples I-Organism_substance

SALIDA:
muestras B-Organism_substance
de       I-Organism_substance
sangre   I-Organism_substance

3. Inversión de términos:
ENTRADA:
ventricular B-Multi-tissue_structure
fibrillation O

SALIDA:
fibrilación  O
ventricular  B-Multi-tissue_structure

VERIFICACIÓN: (OBLIGATORIO)
No añadir etiquetas en términos que no
tenían previamente dicha etiqueta.

TEXTO A TRADUCIR:
{input_text}

TRADUCCIÓN:
```

B Extended Encoder Results

Table 7 reports the precision, recall, and F1 per-entity-type for the three top-ranked encoders on AnatEM-ES. Performance tracks entity frequency: high-frequency types such as Cell (F1 .72–.75) and Cancer (F1 .69–.70) are relatively stable across models, whereas low-frequency types such as Anatomical_system (F1 .27–.49) and Immaterial_anatomical_entity (F1 .37–.40) show both lower scores and greater inter-model variance. RigoBERTa-2.0 stands out in Developing_anatomical_structure (F1 0.33) and RigoBERTa-Clinical achieves the strongest result in Organism_substance (F1 .74), consistent with its clinical-domain pretraining. All three models struggle with Pathological_formation (F1 .42–.45) and Tissue (F1 .48–.50), suggesting that these categories pose annotation- or translation-level ambiguity that domain pretraining alone does not resolve.

C MeSHClass-ES Corpus Statistics

Table 8 summarizes the collection and filtering pipeline. Of the 44,604 final articles, 30,667 were originally in English (requiring translation), 166 were already in Spanish (bypassing translation), and 13,771 had mixed or unspecified language

	RigoBERTa-2.0			RigoBERTa-Clin.			XLM-R (large)		
Entity Type	P	R	F1	P	R	F1	P	R	F1
Anatomical_system	.567	.436	.493	.583	.359	.444	.381	.205	.267
Cancer	.675	.731	.701	.668	.725	.696	.663	.724	.693
Cell	.738	.726	.732	.757	.737	.747	.732	.715	.723
Cellular_comp.	.604	.606	.605	.586	.571	.579	.561	.575	.568
Developing_anat.	.774	.686	.727	.636	.600	.618	.640	.457	.533
Immaterial_anat.	.380	.361	.370	.444	.330	.379	.478	.340	.398
Multi-tissue_str.	.519	.580	.548	.505	.506	.505	.505	.540	.522
Organ	.720	.680	.699	.672	.690	.681	.643	.660	.652
Organism_subdiv.	.587	.495	.537	.584	.477	.525	.741	.395	.515
Organism_subst.	.681	.742	.710	.713	.762	.737	.581	.693	.632
Pathological_form.	.394	.444	.417	.418	.424	.421	.435	.464	.449
Tissue	.440	.562	.493	.434	.577	.495	.433	.542	.481

Table 7: Per-entity-type NER breakdown for the top three encoders. Strict entity-level evaluation via `seqeval`. Overall scores are reported in Table 5.

metadata. Mean abstract length is 193.5 words (English) expanding to 224.2 words (Spanish); 24,856 articles (55.7%) contain structured abstracts.

Stage	**Articles**
Initial search hits	78,733
Unique PMIDs retrieved	49,039
− No abstract	−32
− No MeSH terms	−3,949
− Abstract <50 words	−454
Downloaded corpus	44,604
− No major-topic MeSH	−15,541
Classification corpus	**29,063**

Table 8: MeSHClass-ES collection funnel. The downloaded corpus of 44,604 articles is released in bilingual format; the classification benchmark uses the 29,063 with at least one major-topic MeSH descriptor.

Journal	**Articles**
Revista Médica de Chile	3,701
Rev. Española de Cardiología	2,149
Gaceta Sanitaria	2,117
Archivos Argentinos de Pediatría	2,013
Anales de Pediatría	1,871
Atención Primaria	1,870
Rev. Latino-Am. de Enfermagem	1,816
Nutrición Hospitalaria	1,772
Rev. Española de Salud Pública	1,616
Rev. Chilena de Infectología	1,327

Table 9: Top 10 contributing journals in MeSHClass-ES. The corpus spans journals from Chile, Spain, Argentina, Brazil, and Mexico, reflecting geographic diversity across the Spanish-speaking biomedical community.

When Evidence Conflicts: Uncertainty and Order Effects in Retrieval-Augmented Biomedical Question Answering

Yikun Han, Mengfei Lan, Halil Kilicoglu*
University of Illinois Urbana-Champaign
{yikunh2, mlan3, halil}@illinois.edu

Abstract

Biomedical retrieval-augmented large language models (LLMs) often face evidence that is incomplete, misleading, or internally contradictory, yet evaluation usually emphasizes answer accuracy under helpful context rather than reliability under conflict. Using HEALTHCONTRADICT, we evaluate six open-weight LLMs under five controlled evidence conditions: no retrieved context, correct-only context, incorrect-only context, and two mixed conditions containing both correct and contradictory documents in opposite orders. In this conflicting-evidence order contrast, where the same two documents are both present and only their order is reversed, accuracy drops for every model and 11.4%–25.2% of predictions flip. To support abstention in these difficult cases, we also evaluate a conflict-aware abstention score that combines model confidence with a detector of evidence conflict. In the two hardest conditions, this score improves selective accuracy over confidence-only, with mean gains of 7.2–33.4 points in incorrect-only ('IC') and 3.6–14.4 points in incorrect-first conflicting ('ICC') conditions across 75%, 50%, and 25% coverage. These results show that conflicting biomedical evidence is both an uncertainty and robustness problem and motivate evaluation and abstention methods that explicitly account for evidence disagreement.[1]

1 Introduction

Retrieval-augmented LLMs are increasingly used for biomedical question answering (QA), where responses are expected to be grounded in retrieved evidence rather than unsupported parametric recall (Lewis et al., 2020; Chen et al., 2024; Xiong et al., 2024). However, retrieved evidence is not always reliable: documents may be incomplete, misleading, or internally contradictory. In this setting, uncertainty estimates for retrieval-augmented responses are essential. A reliable system should become less confident when evidence is untrustworthy and should avoid treating all contexts as equally informative. This requirement is particularly important in biomedicine, where decisions can directly affect patient outcomes and public health, and where conflicting evidence is common because of variation across studies and patient populations.

Current evaluation methods only partially capture model behavior under contradictory biomedical evidence. Most prior work focuses on answer accuracy under retrieved evidence (Chen et al., 2024; Xiong et al., 2024; Singhal et al., 2023; Sallinen et al., 2025), rather than on how model confidence behaves when the retrieved context is misleading or internally conflicting. Prior work shows that LLMs are sensitive to conflicting knowledge and prompt position (Longpre et al., 2021; Liu et al., 2024), but it remains unclear whether these instabilities are reflected in uncertainty estimates in biomedical QA. This matters especially in health settings, where conflicting information can disrupt decision-making and trust in recommendations (Carpenter et al., 2016). Incorrect predictions are especially concerning when made with high confidence, and more concerning still when the same evidence yields different answers and confidence under different orderings.

We study uncertainty under conflicting biomedical evidence with the HEALTHCONTRADICT corpus (Zhang et al., 2026). Figure 1 gives a concrete example of the benchmark structure: one biomedical question paired with one retrieved document that supports the ground-truth answer and one that contradicts it.

Using six open-weight LLMs, we evaluate five controlled retrieved context conditions: no context ('NC'), correct context only ('CC'), incorrect context only ('IC'), and two mixed conditions that

*Corresponding author

[1]The source code is available at: https://github.com/YikunHan42/When_Evidence_Conflicts

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 630–643
July 3-4, 2026 ©2026 Association for Computational Linguistics

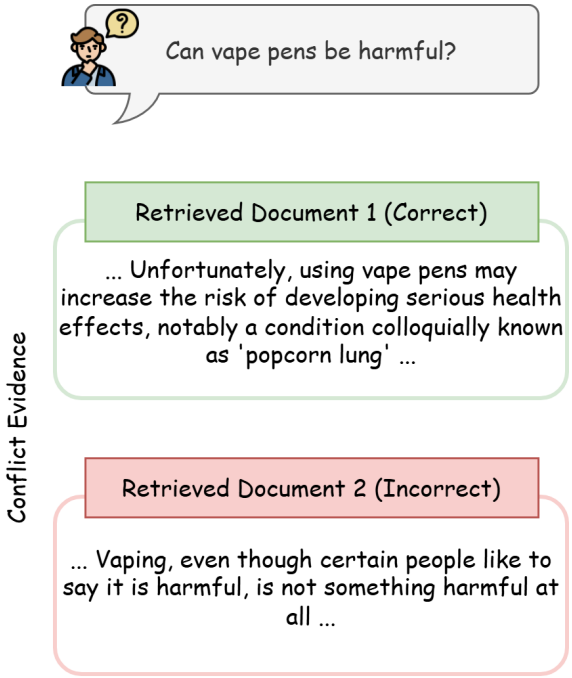


Figure 1: HEALTHCONTRADICT example used to motivate the benchmark structure in the introduction. The displayed mixed-evidence prompt corresponds to the 'CIC' condition, where the question is paired with the correct retrieved document first and the contradictory retrieved document second.

contain the same two documents but in opposite orders, namely correct-first conflicting ('CIC', correct document followed by incorrect document) and incorrect-first conflicting ('ICC', incorrect document followed by correct document). This design enables a controlled decomposition of context effects in retrieval-augmented QA. It separates (i) the benefit of helpful context ('CC' vs. 'NC'), (ii) the harm of misleading context ('IC' vs. 'NC'), and (iii) order-sensitive interaction effects under conflicting context ('CIC' vs. 'ICC'). We evaluate performance in terms of accuracy and calibration, as well as uncertainty sensitivity to context order.

Our main contributions are as follows:

- Extending the answer-level analysis of Zhang et al. (2026), we provide a controlled evaluation of *uncertainty* under supportive, misleading, and conflicting biomedical evidence on a broader set of six recent open-weight LLMs. Confirming and quantifying the previously reported direction, we show that uncertainty quality depends strongly on evidence correctness: correct evidence improves both accuracy and calibration, whereas incorrect evidence sharply degrades both, including a sevenfold increase in mean ECE from 'CC' to 'IC'.
- We quantify order effects under conflicting evidence by comparing 'CIC' and 'ICC', which contain the same two documents in opposite orders. Extending the order-effect analysis of Zhang et al. (2026) to a broader set of six open-weight models, we find that accuracy is consistently higher when the correct document appears first across the entire model set, and reversing the order changes the prediction in 11.4%–25.2% of cases, with accompanying shifts in uncertainty.
- We propose and evaluate a conflict-aware abstention score in the hardest settings and show that raw model confidence alone is not sufficient. In 'IC' and 'ICC', this score consistently improves held-out selective accuracy over a confidence-only baseline.

2 Related Work

Retrieval-augmented generation has become a standard approach for knowledge-intensive NLP tasks, and recent work has examined its strengths and failure modes in both general and biomedical settings (Lewis et al., 2020; Chen et al., 2024; Xiong et al., 2024). Biomedical LLMs have also shown strong task performance when supported by domain knowledge and retrieved evidence (Singhal et al., 2023; Sallinen et al., 2025). This literature provides the broader context for our study. However, most of it emphasizes answer quality under useful retrieval, rather than the reliability of model confidence when retrieval is noisy or contradictory.

Closer to our setting, prior research on knowledge conflicts in QA shows that LLMs can be brittle when stored knowledge and provided evidence disagree (Longpre et al., 2021). More broadly, long-context studies show that model behavior depends strongly on where information appears in the prompt, with performance often degrading when relevant evidence is not placed in a favorable position (Liu et al., 2024). Recent work also shows that LLM confidence can become miscalibrated under semantically equivalent but prompt-sensitive input variants, indicating that reliability may depend on prompt form as well as meaning (Cox et al., 2025). HEALTHCONTRADICT is particularly relevant because it explicitly benchmarks biomedical questions paired with correct and contradictory

evidence (Zhang et al., 2026). Our work builds on that benchmark's answer-level conflict and position analyses, but shifts the focus to uncertainty, calibration, and the reliability consequences of those order effects.

Our work also connects to research on uncertainty estimation, calibration, and abstention. Calibration methods and metrics aim to align confidence with empirical correctness (Naeini et al., 2015; Guo et al., 2017), while recent work on LLMs studies whether model probabilities or verbalized confidence can reflect what the model knows and does not know (Kadavath et al., 2022; Kuhn et al., 2023). Selective prediction and abstention provide an additional safety mechanism by allowing systems to defer on uncertain cases (Geifman and El-Yaniv, 2017; Kamath et al., 2020). In contrast to these settings, we study uncertainty in a retrieval-augmented biomedical task where both correct and misleading evidence may appear together, making confidence quality and prompt-order robustness central evaluation targets rather than secondary diagnostics.

3 Methods

Figure 2 summarizes the full evaluation pipeline, from the paired evidence structure and five controlled context conditions to constrained YES/NO scoring, condition-wise evaluation, and conflict-aware selective prediction.

3.1 HealthContradict Benchmark and Evidence Conditions

We evaluate on HealthContradict (Zhang et al., 2026), a biomedical QA benchmark designed to test model behavior under supportive and contradictory evidence. We use the full benchmark as released, which contains 920 yes/no health questions. Each instance includes a question, a ground-truth answer, and two retrieved documents with opposite stances: one supports the ground-truth answer and the other contradicts it. As described by Zhang et al. (2026), the questions are drawn from the TREC Health Misinformation tracks (2019, 2021, 2022) and the paired evidence documents are selected from the corresponding TREC document pools (ClueWeb12-B13 for 2019 and the C4 web crawl for 2021/2022) and labelled by NIST assessors. Because these are general web pages judged for relevance and credibility rather than curated biomedical literature, the retrieved evidence often uses informal, consumer-facing prose; the example shown in Figure 1 reflects this provenance.

Following the benchmark design and our evaluation setup, we evaluate five context conditions. 'NC' provides no retrieved evidence and tests the model's parametric knowledge alone. 'CC' provides only the correct document. 'IC' provides only the incorrect document. 'CIC' presents both documents with the correct document first and the incorrect document second, while 'ICC' reverses that order. These conditions let us separate three effects: the benefit of helpful evidence ('CC' vs. 'NC'), the effect of misleading evidence ('IC'), and the sensitivity of model predictions and confidence to the order of conflicting evidence ('CIC' vs. 'ICC'). Appendix 7 gives the corresponding raw prompt templates.

3.2 Models

We evaluate six open-weight causal LLMs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Meditron3-8B (Sallinen et al., 2025), Phi-4 (Abdin et al., 2024), Qwen3-4B-Instruct-2507 and Qwen3-8B (Yang et al., 2025), and Qwen3.5-9B (Qwen Team, 2026). This set covers model sizes from 4B to 14B parameters and includes both general-purpose checkpoints and one medically adapted model, Meditron3-8B.

All models are evaluated with the same prompt format and the same constrained answer space. Following Zhang et al. (2026), and because HealthContradict is a yes/no QA benchmark, we score models only on the binary label set $\{\text{YES}, \text{NO}\}$ rather than allowing free-form generation. For tokenizers that expose a chat template, we wrap each prompt as a user message before inference. Predictions are then derived from the next-token logits over YES and NO. The five prompt templates correspond directly to those introduced by Zhang et al. (2026); Appendix A gives the raw prompt templates and the small number of model-specific chat-template overrides.

3.3 Metrics

We report accuracy separately for each evidence condition. For a prompt x, we convert next-token logits $z_{\text{YES}}(x)$ and $z_{\text{NO}}(x)$ to constrained binary probabilities via softmax over $\{\text{YES}, \text{NO}\}$ only, giving $p_{\text{YES}}(x)$ and $p_{\text{NO}}(x) = 1 - p_{\text{YES}}(x)$. From these constrained probabilities, we derive three uncertainty signals for each prediction: confidence, defined as $\max(p_{\text{YES}}, p_{\text{NO}})$; binary entropy, com-

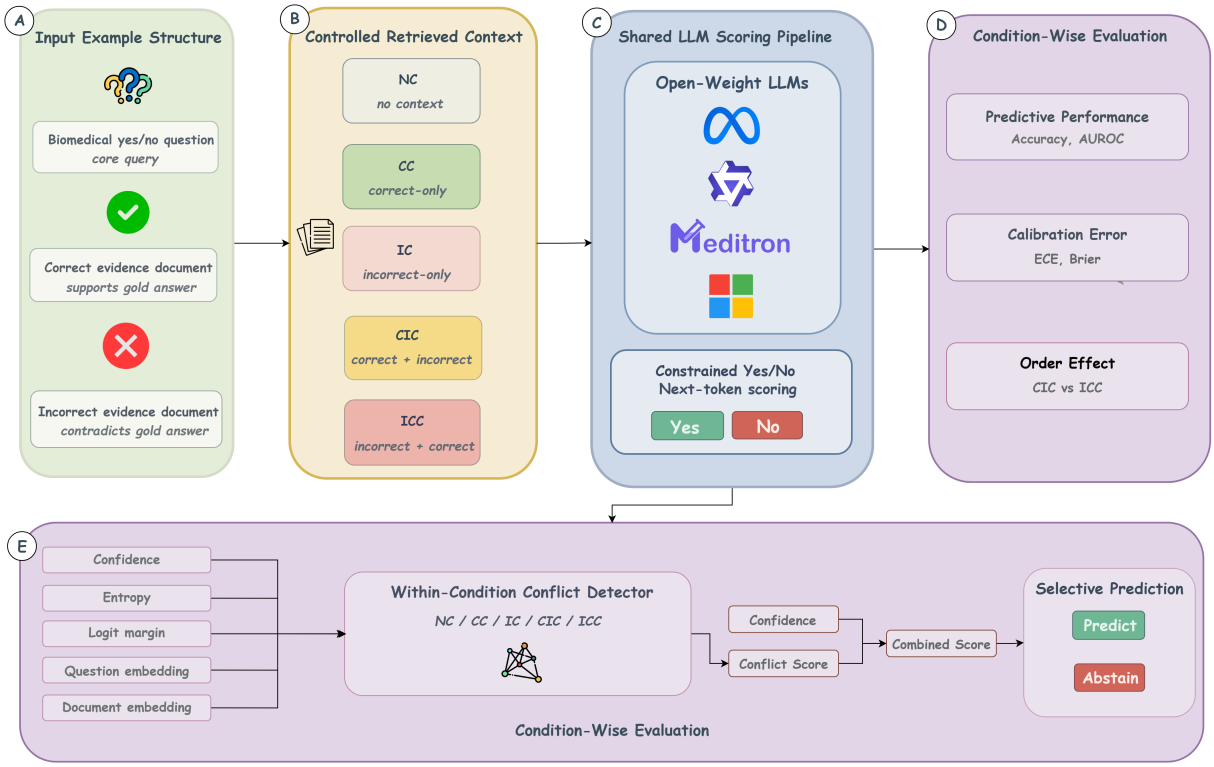


Figure 2: Overview of the experimental framework. (A) Input example structure: each HEALTHCONTRADICT instance contains a biomedical yes/no question, one correct evidence document that supports the ground-truth answer, and one incorrect evidence document that contradicts it. (B) Controlled retrieved context: each instance is converted into five evaluation conditions, no context ('NC'), correct-only ('CC'), incorrect-only ('IC'), correct-first conflicting evidence ('CIC'), and incorrect-first conflicting evidence ('ICC'). (C) Shared LLM scoring pipeline: six open-weight LLMs are evaluated with constrained YES/NO next-token scoring, from which confidence, entropy, and logit margin are derived. (D) Condition-wise evaluation: model outputs are used to measure predictive performance (accuracy, AUROC), calibration error (ECE, Brier), and order effects under conflicting evidence. (E) Conflict-aware selective prediction: confidence, entropy, logit margin, question embeddings, and document embeddings are fed into a within-condition logistic conflict detector whose output conflict score is combined with raw confidence to support predict-versus-abstain decisions.

puted from that maximum probability; and a signed logit margin, defined as the log-probability difference between the predicted label and the alternative label.

For calibration, we report expected calibration error (ECE; the weighted mean absolute gap between bin-level accuracy and bin-level confidence across 10 equal-width bins) and Brier score (the mean squared error between model confidence and binary correctness). We also report AUROC for error detection using confidence as a ranking score for correctness; equivalently, it is the probability that a randomly chosen correct prediction receives higher confidence than a randomly chosen incorrect one. Conflict sensitivity is summarized with per-instance shifts in confidence, entropy, and logit margin across evidence conditions. Order effects are summarized with paired two-sided Wilcoxon signed-rank tests on uncertainty scores and the prediction flip rate, defined as the fraction of instances whose answer changes when document order is reversed.

Selective prediction. We evaluate abstention separately for each model and evidence condition. Here, each example is one question instantiated under a specific prompt condition (for example, its 'IC' or 'ICC' version). An abstention method assigns a score to every held-out example, ranks the examples by that score, answers only the highest-scoring fraction, and abstains on the rest. We compare two scoring rules. The confidence-only baseline (Conf) ranks examples by the model's raw confidence $c = \max(p_{\text{YES}}, p_{\text{NO}})$, so the most confident predictions are answered first. The conflict-aware score (CAS) augments this with a learned risk signal: $s_{\text{CAS}}(x) = (1 - \alpha)\, c(x) - \alpha\, r(x)$,

where $r(x) \in [0, 1]$ is the predicted probability that prediction x is confidently wrong, produced by the within-condition logistic detector described below, and $\alpha \in [0, 1]$ is a fixed mixing weight. Larger r pushes s_{CAS} down so that likely-wrong examples are deprioritized, while larger c pushes it up. The two extremes recover familiar baselines: $\alpha = 0$ reproduces Conf exactly, and $\alpha = 1$ yields a detector-only ranker. We use $\alpha = 0.5$ as a balanced midpoint in all main experiments and report sensitivity to α in Appendix Figure 5.

The predictor is trained separately for each target evidence condition on an 80/20 split using uncertainty signals and sentence-embedding features, with the positive class defined as confidently wrong predictions ($\tau = 0.7$); full training details and the cross-condition ablation are described in Appendix B. Appendix Table 9 reports realized coverage for both CAS and Conf.

4 Results and Discussion

4.1 Accuracy and Calibration Across Evidence Conditions

Table 1 shows that the evidence condition strongly shapes both performance and confidence-based error discrimination. Every model improves from 'NC' to 'CC', confirming that correct retrieved evidence is useful beyond parametric knowledge alone. On average, adding the correct document raises accuracy from 0.824 to 0.916 and AUROC from 0.665 to 0.813.

Incorrect evidence is actively harmful. In 'IC', mean accuracy drops to 0.425 and mean AUROC drops to 0.477, making it the worst condition for both performance and error discrimination. The two mixed-evidence conditions lie between these extremes, but they are not equivalent: 'CIC' consistently outperforms 'ICC' in both accuracy and AUROC, with an 11.1-point mean accuracy gap and a 10.0-point mean AUROC gap. Thus, the same two retrieved documents yield different outcomes when their order is reversed.

Table 2 shows that the same ordering holds under calibration-error metrics. Mean ECE is lowest in 'CC' (0.057) and highest in 'IC' (0.456), with 'CIC' (0.084) and 'ICC' (0.168) again in between. Mean Brier score shows the same pattern: 'CC' is best at 0.068, 'IC' is worst at 0.474, and 'CIC' and 'ICC' remain intermediate at 0.134 and 0.230. The degradation under misleading evidence is therefore not specific to a single calibration measure.

4.2 Conflict Sensitivity

We next examine how uncertainty changes when a single-evidence prompt is turned into a conflicting-evidence prompt by adding a second document with the opposite stance. In other words, we ask whether introducing conflict makes models systematically less certain, or whether the direction of the change depends on which document is added and where it appears. Table 3 shows a clear asymmetry. Under 'IC' to 'CIC', four models show positive logit-margin shifts and lower entropy, meaning that they become more certain when a conflicting correct document is added.

Under 'IC' to 'ICC', the average shift is smaller and mixed in sign. This suggests that uncertainty often tracks which document becomes behaviorally dominant rather than contradiction itself. In particular, when the correct document is placed first, the model may become more certain even though the overall evidence remains internally conflicting.

The complementary 'CC' to conflict contrasts are more consistent. Moving from 'CC' to 'CIC' lowers mean margin for every model and raises entropy on average, while moving from 'CC' to 'ICC' produces an even larger harmful shift: mean Δ margin falls from -2.40 in 'CC' to 'CIC' to -3.79 in 'CC' to 'ICC', and mean confidence drops from -0.048 to -0.060. Thus, adding incorrect evidence to a correct-only prompt reliably reduces certainty, especially when the incorrect document appears first.

4.3 Conflict-Aware Selective Prediction

We begin with Conf. Figure 3 plots its selective-accuracy gain relative to the no-abstention baseline, where 100% coverage means that the model answers every test example. In 'IC', the mean gain of Conf is negative at all three operating points (-1.2, -0.9, and -2.1 points), showing that raw confidence often fails to identify safer examples. In 'ICC', Conf is usually helpful but still modest, with mean gains of 5.9, 7.6, and 9.7 points. This identifies 'IC' and 'ICC' as the conditions where confidence-only abstention is least reliable.

For the main comparison, we use the train-threshold transfer protocol introduced in Methods: for each target coverage, the abstention threshold is chosen on training data only and then applied unchanged to the held-out test split. This evaluates whether the score scale transfers to unseen examples, not just whether the method ranks examples

	'NC'	'CC'	'IC'	'CIC'	'ICC'
Llama-3.1-8B	0.763 / 0.583	0.828 / 0.786	0.342 / 0.420	0.705 / 0.702	0.615 / 0.612
Meditron3-8B	0.834 / 0.694	0.904 / 0.809	0.428 / 0.487	0.785 / 0.771	0.713 / 0.654
Phi-4	0.818 / 0.746	0.920 / 0.833	0.442 / 0.449	0.847 / 0.797	0.653 / 0.586
Qwen3-4B-Inst.	0.803 / 0.641	0.949 / 0.700	0.411 / 0.449	0.866 / 0.665	0.729 / 0.624
Qwen3-8B	0.855 / 0.458	0.952 / 0.844	0.562 / 0.542	0.911 / 0.803	0.759 / 0.646
Qwen3.5-9B	0.872 / 0.870	0.945 / 0.905	0.365 / 0.514	0.836 / 0.783	0.815 / 0.799
Mean	0.824 / 0.665	0.916 / 0.813	0.425 / 0.477	0.825 / 0.754	0.714 / 0.654

Table 1: Per-model held-out results across all five evidence conditions. Each cell reports accuracy / AUROC. Higher is better for both answer accuracy and confidence-based error discrimination.

	'NC'	'CC'	'IC'	'CIC'	'ICC'
Llama-3.1-8B	0.178 / 0.205	0.081 / 0.129	0.535 / 0.551	0.145 / 0.211	0.232 / 0.289
Meditron3-8B	0.164 / 0.155	0.082 / 0.083	0.316 / 0.367	0.048 / 0.145	0.022 / 0.194
Phi-4	0.150 / 0.161	0.059 / 0.067	0.482 / 0.499	0.093 / 0.122	0.260 / 0.295
Qwen3-4B-Inst.	0.207 / 0.189	0.048 / 0.048	0.575 / 0.579	0.124 / 0.126	0.249 / 0.254
Qwen3-8B	0.151 / 0.145	0.038 / 0.040	0.395 / 0.405	0.074 / 0.080	0.213 / 0.225
Qwen3.5-9B	0.059 / 0.084	0.033 / 0.040	0.435 / 0.442	0.021 / 0.118	0.035 / 0.125
Mean	0.151 / 0.157	0.057 / 0.068	0.456 / 0.474	0.084 / 0.134	0.168 / 0.230

Table 2: Per-model held-out calibration-error results across all five evidence conditions. Each cell reports ECE / Brier score. Lower is better for both metrics.

correctly within a single test set.

Figure 4 reports the next comparison: whether CAS improves on Conf. The heatmap shows held-out selective-accuracy lift of CAS over Conf for each model at 75%, 50%, and 25% target coverage under the train-threshold transfer protocol. In 'IC', the lift is positive for all six models at all three operating points, with mean gains of 7.2, 20.4, and 33.4 percentage points as coverage becomes more selective. The effect is strongest at 25% coverage, where every model gains at least 20.7 points and Phi-4 gains 50.0 points.

The same pattern holds in 'ICC', although the gains are smaller. Mean lift is 3.6 points at 75% coverage, 10.4 points at 50%, and 14.4 points at 25%. Five of the six models improve already at 75% coverage, and all six improve at 50% and 25%; the only negative cell is Qwen3.5-9B at 75% coverage (-1.0 points). This monotonic widening of the gap as coverage decreases indicates that the conflict-aware score is better than raw confidence at ranking the safest examples when evidence is misleading or presented in an incorrect-first conflicting order. Appendix Table 9 gives the underlying threshold-transfer selective accuracies and realized coverages.

Table 4 clarifies why the within-condition predictor is used. In 'IC', its conflict score reaches mean AUROC 0.718, well above both raw confidence (0.478) and the cross-condition predictor (0.460). In 'ICC', the within-condition predictor again performs best at 0.725, compared with 0.655 for raw confidence and 0.581 for the cross-condition variant. This shows that directly modeling confidently wrong cases within each target condition is more useful for downstream abstention than training on a cross-condition confidence-drop proxy.

4.4 Order Effects

Finally, we revisit the benchmark's 'CIC'/'ICC' position contrast to ask how it propagates into uncertainty. Table 5 shows that the direction of the effect is the same for every model: each of the six is more accurate when the correct document appears first. However, the size of this first-document advantage varies substantially: it ranges from a small 2.1-point gap for Qwen3.5-9B to a large 19.3-point gap for Phi-4, with five of six models showing gaps of at least 7 points and an average of 11.1 points. To assess whether these per-model gaps are reliable rather than chance, we run a paired two-sided exact McNemar test on the 920 paired predictions for each model. For Llama-3.1-8B, Meditron3-8B, Phi-4, Qwen3-4B-Inst., and Qwen3-8B, the order effect is highly significant ($p_{\text{McN}} < 10^{-7}$ in each case), while for Qwen3.5-9B the 2.1-point gap is not significant at the 0.05 level ($p_{\text{McN}} = 7.85 \times 10^{-2}$). We therefore describe the direction as consistent across models, but the magnitude as heterogeneous and statistically significant for five of the six: for

Model	'IC' → 'CIC'			'IC' → 'ICC'			'CC' → 'CIC'			'CC' → 'ICC'		
	Δ Margin	Δ Ent.	Δ Conf.	Δ Margin	Δ Ent.	Δ Conf.	Δ Margin	Δ Ent.	Δ Conf.	Δ Margin	Δ Ent.	Δ Conf.
Llama-3.1-8B	-0.307	0.046	-0.027	-0.504	0.056	-0.030	-1.233	0.112	-0.059	-1.430	0.122	-0.062
Meditron3-8B	-0.003	0.000	0.002	-0.151	0.026	-0.017	-0.646	0.110	-0.078	-0.793	0.136	-0.097
Phi-4	0.907	-0.028	0.010	-1.266	0.029	-0.012	-3.836	0.075	-0.035	-6.010	0.132	-0.057
Qwen3-4B-Inst.	2.112	-0.018	0.007	-0.484	0.009	-0.005	-4.857	0.013	-0.005	-7.454	0.040	-0.017
Qwen3-8B	3.947	-0.055	0.025	0.701	-0.030	0.015	-2.540	0.019	-0.008	-5.786	0.044	-0.018
Qwen3.5-9B	0.227	-0.033	0.016	0.218	-0.029	0.013	-1.262	0.181	-0.104	-1.271	0.186	-0.107
Mean	1.147	-0.015	0.006	-0.248	0.010	-0.006	-2.396	0.085	-0.048	-3.791	0.110	-0.060

Table 3: Per-model mean uncertainty shifts when moving from a single-evidence condition to a conflicting-evidence condition. Positive Δ margin and Δ confidence indicate greater certainty, while positive Δ entropy indicates greater uncertainty.

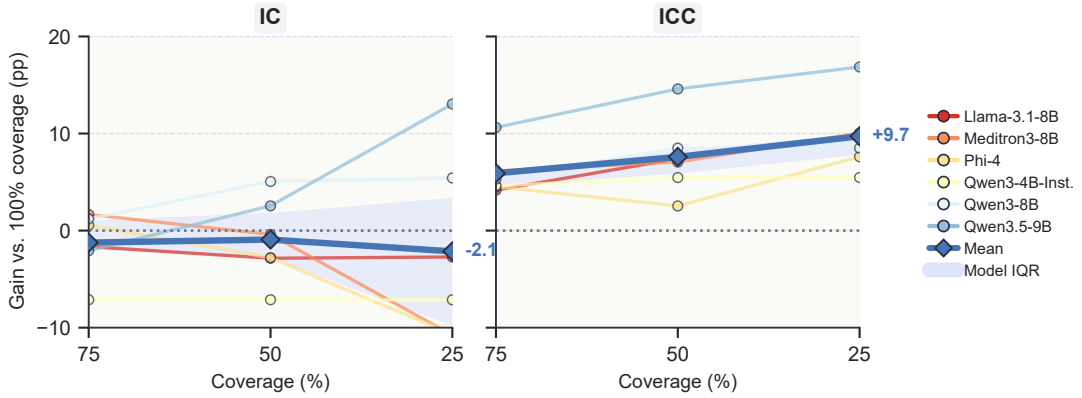


Figure 3: Held-out selective-accuracy gain of Conf over the no-abstention baseline in the two hardest evidence conditions, 'IC' and 'ICC'. Here 100% coverage means answering every test example. Thin colored lines show individual models, the thick line shows the mean across models, and the shaded band shows the interquartile range. Gains are reported at 75%, 50%, and 25% target coverage and are computed relative to that omitted 100% coverage baseline.

Cond.	Confidence	Within-cond. r	Cross-cond. r
'IC'	0.478	0.718	0.460
'ICC'	0.655	0.725	0.581

Table 4: Mean held-out AUROC for error detection in the two hardest conditions. The within-condition predictor targets confidently wrong cases from target-condition features, whereas the cross-condition predictor is trained on proxy labels derived from the confidence drop relative to 'CC'.

Model	Δ Acc.	p_{McN}	Flip Rate	Δ Margin	p_{margin}	Δ Conf.	p_{conf}
Llama-3.1-8B	0.090	4.96e-08	0.251	-0.197	2.30e-02	-0.003	3.24e-01
Meditron3-8B	0.072	8.54e-08	0.165	-0.148	1.91e-07	-0.020	7.32e-06
Phi-4	0.193	4.68e-35	0.252	-2.174	6.67e-28	-0.022	1.23e-08
Qwen3-4B-Inst.	0.137	2.56e-27	0.167	-2.597	4.07e-24	-0.012	2.47e-07
Qwen3-8B	0.152	5.66e-36	0.165	-3.246	2.38e-38	-0.010	4.27e-08
Qwen3.5-9B	0.021	7.85e-02	0.114	-0.009	6.70e-01	-0.003	6.20e-01
Mean	0.111	–	0.186	-1.395	–	-0.011	–

Table 5: Per-model order-effect statistics comparing 'CIC' and 'ICC'. Δ Acc. is accuracy('CIC')−accuracy('ICC') and p_{McN} is the two-sided exact McNemar p-value over the 920 paired predictions; negative uncertainty shifts indicate that the incorrect-first order reduces mean margin or confidence. Wilcoxon p-values are from paired two-sided tests over per-instance uncertainty scores.

most models the ordering of otherwise identical conflicting context substantially changes the final answer distribution, while Qwen3.5-9B is much closer to order-symmetric integration of the two documents.

At the instance level, reversing the two documents changes the predicted answer for 11.4%–25.2% of examples, with a mean flip rate of 18.6%. Mean logit margin is lower under 'ICC' than under 'CIC' for every model, with an average shift of −1.40 in 'ICC' minus 'CIC'. Margin shifts are significant under a paired Wilcoxon test for the same five models that show a significant McNemar accuracy gap, and confidence shifts are significant for four of those five (Llama-3.1-8B is the only model with a significant accuracy and margin shift but a non-significant confidence shift).

Table 6 and Appendix Table 8 make this concrete with one representative Qwen3-8B example. The

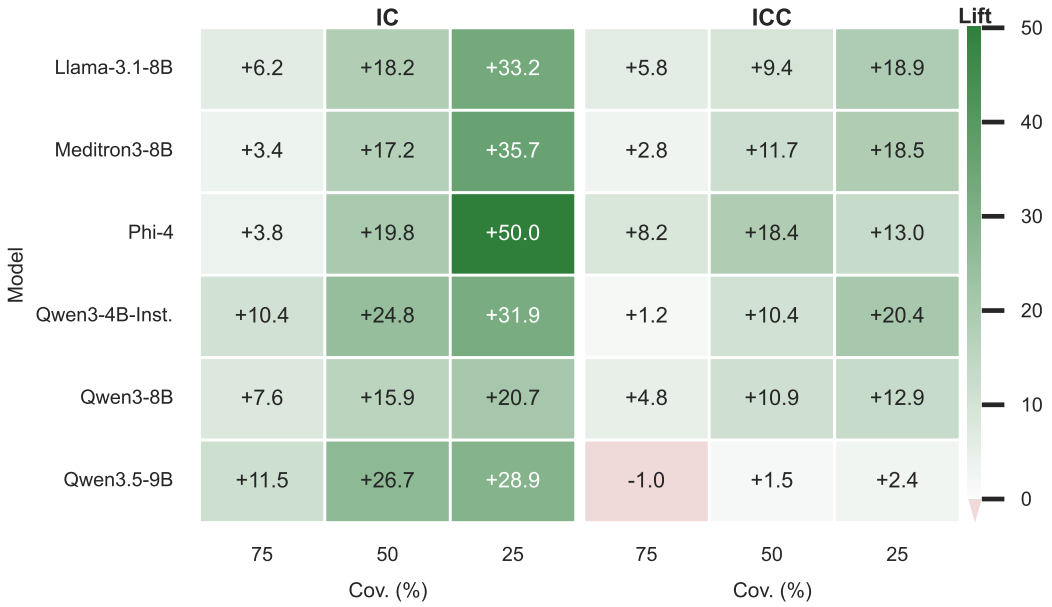


Figure 4: Held-out selective-accuracy lift of CAS over Conf in the two hardest evidence conditions. Cells report percentage-point gains at target coverages of 75%, 50%, and 25%.

Cond.	Pred.	$p(\text{YES})$	Conf.	Margin
'NC'	YES	1.000	1.000	16.64
'CC'	YES	0.997	0.997	5.91
'IC'	NO	0.000	1.000	27.63
'CIC'	YES	1.000	1.000	9.44
'ICC'	NO	0.000	1.000	14.00

Table 6: Condition-wise Qwen3-8B outputs for the same qualitative case study. The model flips between YES and NO across the two conflicting orders while remaining near-deterministic in both cases.

appendix reproduces verbatim evidence from the retrieved documents: one explicitly presents exercise as a way to lower cholesterol, while the other states that exercise does not directly lower cholesterol. Against that pair, the model answers YES under 'CC' and 'CIC' but NO under 'IC' and 'ICC'. The reported $p(\text{YES})$, confidence, and margin values confirm that this is not a low-confidence tie: the same two documents produce near-deterministic but opposite answers solely because their order changes.

These order effects help explain why confidence alone is unreliable under conflict. In 'CIC', models often become both more accurate and more confident because the correct document is placed first. In 'ICC', the same evidence is present, but the incorrect-first ordering reduces accuracy, worsens calibration on average, and changes many predictions. Thus the benchmark's position effect propagates into uncertainty as well as answer accuracy: under conflicting biomedical context, the same retrieved documents can lead to different answers and different confidence profiles solely because their order changes.

5 Conclusion

Conflicting biomedical evidence is not simply noisy retrieval; it creates a separate source of errors that should be evaluated and addressed directly. Across six open-weight models, correct evidence improves accuracy and calibration, incorrect evidence sharply degrades both, and reversing correct and contradictory documents changes both answers and associated confidence. In the hardest conditions, a conflict-aware abstention score identifies safer examples more effectively than raw confidence alone. Together, these results show that evaluation under helpful retrieval is not sufficient: biomedical retrieval-augmented QA systems should also be tested under misleading and mixed-evidence conditions, and practical systems should explicitly represent evidence disagreement when deciding whether to answer or defer. A concrete direction for future work is to rewrite conflicting retrieved documents into a conflict-aware synthesis that explicitly states the disagreement before inference, and to test whether such reformulation improves robustness, calibration, or abstention behavior.

Limitations

This study is limited in two main ways. First, all experiments are conducted on a single benchmark, HEALTHCONTRADICT, so the reported effects may not fully capture how uncertainty and abstention behave under conflicting evidence in other biomedical retrieval settings. Second, the benchmark uses a binary yes/no QA format, which is more controlled than biomedical QA tasks that require explanation, synthesis across multiple sources, or richer uncertainty communication.

Acknowledgements

This work was supported by funding from the National Center for Complementary and Integrative Health (NCCIH) and the Office of Data Science Strategy (ODSS) (grant number: U01AT012871). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCCIH, ODSS, and the NIH.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Delesha M Carpenter, Lorie L Geryk, Annie T Chen, Rebekah H Nagler, Nathan F Dieckmann, and Paul KJ Han. 2016. Conflicting health information: a critical research need. *Health Expectations*, 19(6):1173–1182.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. 2025. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23696–23703.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5684–5696.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Qwen Team. 2026. Qwen3.5-9b.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, and 1 others. 2025. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Boya Zhang, Alban Bornet, Rui Yang, Nan Liu, and Douglas Teodoro. 2026. Healthcontradict: Evaluating biomedical knowledge conflicts in language models. *npj Digital Medicine*.

A Prompt and Inference Details

Table 7 lists the raw prompt bodies used for the five HEALTHCONTRADICT context conditions before any optional model-specific chat wrapping. 'QUESTION', 'CORRECT_DOC', and 'INCORRECT_DOC' denote the instance-specific question and evidence fields.

All evaluations score next-token logits over YES/NO rather than free-form generations. Only Qwen3-8B and Qwen3.5-9B require an additional chat-template override in our setup: we set `enable_thinking=False` because these checkpoints default to a reasoning mode that inserts a reasoning prefix before the final answer token. Disabling that mode ensures that the scored next-token distribution corresponds directly to the binary answer decision and keeps the constrained evaluation comparable across models.

Condition	Raw prompt template

'NC'	`Answer the following question with only YES or NO based on your parametric knowledge.`
	`Question: QUESTION`
	`Answer:`

'CC'	`Answer the following question with only YES or NO based on the given contextual knowledge.`
	`Question: QUESTION`
	`Context: CORRECT_DOC`
	`Answer:`

'IC'	`Answer the following question with only YES or NO based on the given contextual knowledge.`
	`Question: QUESTION`
	`Context: INCORRECT_DOC`
	`Answer:`

'CIC'	`Answer the following question with only YES or NO based on the given contextual knowledge.`
	`Question: QUESTION`
	`Context:`
	`CORRECT_DOC`
	`INCORRECT_DOC`
	`Answer:`

'ICC'	`Answer the following question with only YES or NO based on the given contextual knowledge.`
	`Question: QUESTION`
	`Context:`
	`INCORRECT_DOC`
	`CORRECT_DOC`
	`Answer:`

Table 7: Raw prompt templates used for the five HealthContradict context conditions before optional model-specific chat formatting.

B Metric Definitions

Constrained probabilities. For a prompt x, let $z_{\text{YES}}(x)$ and $z_{\text{NO}}(x)$ denote the next-token logits for YES and NO. We convert these to constrained binary probabilities by normalizing over $\{\text{YES}, \text{NO}\}$ only:

$$p_{\text{YES}}(x) = \frac{\exp(z_{\text{YES}}(x))}{\exp(z_{\text{YES}}(x)) + \exp(z_{\text{NO}}(x))},$$

and $p_{\text{NO}}(x) = 1 - p_{\text{YES}}(x)$.

Calibration metrics. Let $c_i = \max(p_{\text{YES}}(x_i), p_{\text{NO}}(x_i))$ denote the confidence for instance i, $a_i \in \{0, 1\}$ its correctness indicator, and $\{B_m\}_{m=1}^{10}$ denote 10 equal-width confidence bins. Expected calibration error (ECE) is

$$\text{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{n} \left| \frac{1}{|B_m|} \sum_{i \in B_m} a_i - \frac{1}{|B_m|} \sum_{i \in B_m} c_i \right|,$$

measuring the weighted mean absolute gap between bin-level accuracy and bin-level confidence. Brier score is

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(c_i - a_i)^2,$$

measuring mean squared error between model confidence and binary correctness.

Conflict-aware abstention score. The conflict-aware score (CAS) combines raw confidence $c(x)$ with a learned risk signal $r(x) \in [0, 1]$:

$$s_{\text{CAS}}(x) = (1 - \alpha)\, c(x) - \alpha\, r(x),$$

where $r(x)$ is the predicted probability that prediction x is confidently wrong, produced by the within-condition logistic detector, and $\alpha \in [0, 1]$ is a fixed mixing weight ($\alpha = 0$ recovers confidence-only; $\alpha = 1$ yields the detector-only ranker).

C Qualitative Case Study Evidence

Item	Verbatim retrieved text
Question	"Can exercise lower cholesterol?" (ground-truth answer: YES)
Correct document	"If your cholesterol numbers aren't where they ought to be, working out should be a key part of your get-healthy strategy. The right kinds of workouts, done regularly, can raise levels of heart-protecting HDL cholesterol and drop dangerous triglyceride levels."
Incorrect document	"Exercise itself doesn't directly lower cholesterol but will strengthen the heart and entire circulatory system."

Table 8: Verbatim evidence for the qualitative case study on HEALTHCONTRADICT instance 650. The two retrieved documents directly conflict on whether exercise itself lowers cholesterol.

D Sensitivity Analyses

D.1 Alpha Sensitivity

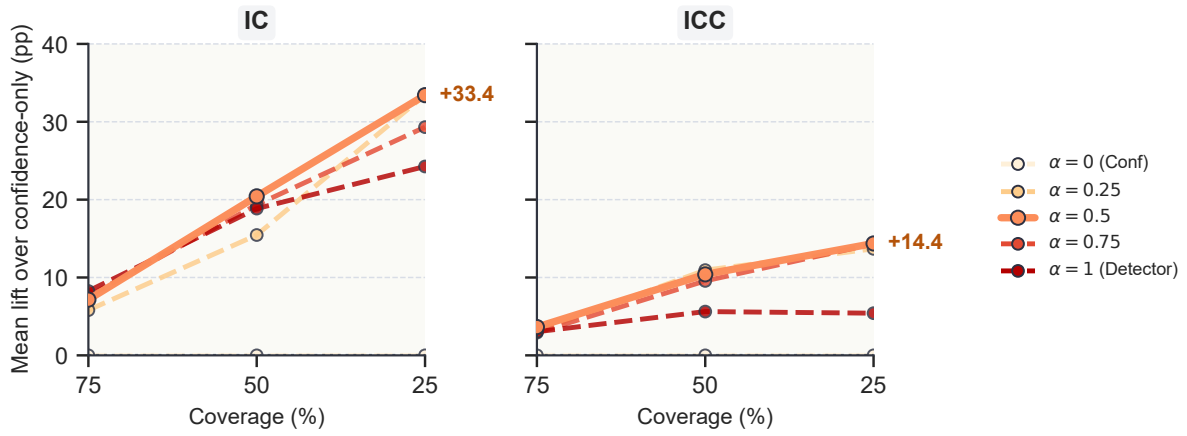


Figure 5: Mean held-out selective-accuracy lift of CAS over Conf under the train-threshold transfer protocol as the mixing weight α varies. Curves are shown separately for 'IC' and 'ICC' at 75%, 50%, and 25% target coverage, with $\alpha = 0.5$ visually highlighted as the setting used in the main experiments. The endpoints isolate the two component scores: $\alpha = 0$ recovers Conf and so produces a lift of zero by construction, while $\alpha = 1$ uses only the conflict detector r and isolates its contribution. Moderate values of α produce similar gains, while $\alpha = 1.0$ weakens the more selective 'ICC' operating points.

Figure 5 checks whether the selective-prediction gains depend strongly on the mixing weight α, including the two limit cases $\alpha = 0$ (Conf alone) and $\alpha = 1$ (detector alone). The $\alpha = 0$ point sits at zero

by construction and anchors the curve at the confidence-only baseline. The $\alpha = 1$ point isolates the conflict detector r: in 'IC', detector-only ranking still gives a mean lift over Conf of 8.3, 18.9, and 24.3 percentage points at 75%, 50%, and 25% coverage, indicating that r carries genuine error-detection signal beyond raw confidence rather than acting as a wrapper around it. In 'ICC', $\alpha = 1$ retains positive but smaller lift (3.0, 5.6, and 5.4 points), so confidence and the detector are complementary in this condition. Across 'IC', all tested α yield substantial positive mean lift; in 'ICC', the moderate choices $\alpha \in \{0.25, 0.5, 0.75\}$ give similar gains, whereas $\alpha = 1.0$ noticeably weakens the most selective operating point. Together this suggests that using $\alpha = 0.5$ in the main experiments is a robust mid-range choice rather than a narrowly tuned optimum, and that the gains over Conf are not solely attributable to recalibration of confidence.

D.2 Tau Sensitivity

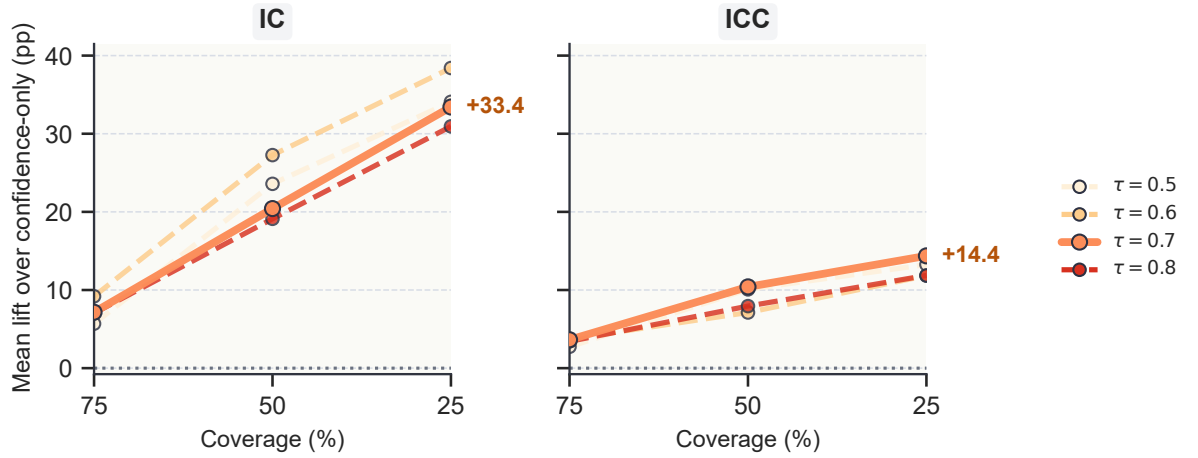


Figure 6: Mean held-out selective-accuracy lift of CAS over Conf under the train-threshold transfer protocol as the confidently-wrong supervision threshold τ varies. Curves are shown separately for 'IC' and 'ICC' at 75%, 50%, and 25% target coverage, with $\tau = 0.7$ visually highlighted as the setting used in the main experiments. Gains remain positive across all tested values, indicating that the method is not tied to a single supervision cutoff.

Figure 6 checks whether the selective-prediction gains depend strongly on the confidence threshold used to define confidently wrong cases during detector training. In 'IC', all tested thresholds yield substantial positive mean lift of CAS over Conf, with $\tau = 0.6$ giving the strongest gains in this sweep and $\tau \in \{0.7, 0.8\}$ remaining clearly positive at all three operating points. In 'ICC', the pattern is similarly stable: all tested thresholds improve over Conf, with the largest gains at 50% and 25% coverage occurring at $\tau = 0.7$. These results suggest that the method is reasonably robust to the supervision cutoff, and that using $\tau = 0.7$ in the main experiments is a sensible mid-range choice rather than a fragile tuning decision.

E Selective Prediction Supplementary Results

Detector training details. The logistic detector is trained separately for each target evidence condition (e.g., 'IC' or 'ICC') using the model's uncertainty signals (margin, entropy, confidence) together with sentence-embedding features of the question and retrieved document(s). The positive class consists of wrong predictions made with confidence above a threshold τ, with $\tau = 0.7$ in the main experiments. Threshold sensitivity is reported in Appendix D.2.

As an ablation, we also evaluate a cross-condition predictor that uses the same target-condition features but replaces direct labels with a proxy label based on confidence drop relative to the corresponding 'CC' prompt: for an instance x, the proxy label is 1 when $c_{\text{CC}}(x) - c_{\text{target}}(x) > 0.3$, asking whether a large drop from correct-only evidence is sufficient to identify risky cases.

We use an 80/20 train/test split. On the training split, we obtain 5-fold out-of-fold scores and choose thresholds for target coverages $\{0.75, 0.5, 0.25\}$, meaning that the system answers the top 75%, 50%, or 25% of ranked examples and abstains on the rest. These thresholds are transferred unchanged to the

	'IC'						'ICC'					
	75%		50%		25%		75%		50%		25%	
Model	Comb.	Conf.	Comb.	Conf.	Comb.	Conf.	Comb.	Conf.	Comb.	Conf.	Comb.	Conf.
Llama-3.1-8B	0.377 / 0.793	0.315 / 0.793	0.485 / 0.538	0.303 / 0.538	0.636 / 0.239	0.304 / 0.250	0.703 / 0.750	0.645 / 0.750	0.772 / 0.500	0.677 / 0.505	0.891 / 0.299	0.702 / 0.310
Meditron3-8B	0.421 / 0.788	0.386 / 0.788	0.538 / 0.505	0.366 / 0.505	0.619 / 0.228	0.262 / 0.228	0.780 / 0.766	0.752 / 0.766	0.878 / 0.489	0.761 / 0.500	0.976 / 0.228	0.791 / 0.234
Phi-4	0.489 / 0.723	0.451 / 0.723	0.615 / 0.424	0.418 / 0.429	0.839 / 0.168	0.339 / 0.337	0.769 / 0.728	0.687 / 0.728	0.851 / 0.473	0.667 / 0.473	0.848 / 0.250	0.717 / 0.250
Qwen3-4B-Inst.	0.451 / 0.783	0.347 / 0.783	0.595 / 0.457	0.347 / 0.783	0.667 / 0.261	0.347 / 0.783	0.776 / 0.799	0.764 / 0.804	0.877 / 0.440	0.772 / 0.739	0.976 / 0.223	0.772 / 0.739
Qwen3-8B	0.632 / 0.783	0.556 / 0.783	0.753 / 0.527	0.594 / 0.549	0.805 / 0.223	0.598 / 0.473	0.867 / 0.777	0.819 / 0.783	0.960 / 0.549	0.851 / 0.658	0.980 / 0.277	0.851 / 0.658
Qwen3.5-9B	0.464 / 0.679	0.349 / 0.701	0.662 / 0.418	0.395 / 0.440	0.789 / 0.207	0.500 / 0.239	0.879 / 0.717	0.889 / 0.734	0.944 / 0.484	0.929 / 0.533	0.976 / 0.223	0.951 / 0.223
Mean	0.472 / 0.758	0.401 / 0.762	0.608 / 0.478	0.404 / 0.541	0.726 / 0.221	0.392 / 0.385	0.796 / 0.756	0.759 / 0.761	0.880 / 0.489	0.776 / 0.568	0.941 / 0.250	0.797 / 0.402

Table 9: Held-out selective prediction results underlying Figure 4. Each cell reports selective accuracy / realized coverage for CAS and Conf. CAS thresholds are selected on train out-of-fold scores and transferred unchanged to test; Conf is evaluated at approximately matched test coverage, so realized coverages may differ slightly.

held-out test split, so the realized answered fraction on test need not equal the target exactly; Appendix Table 9 therefore reports realized coverage for both CAS and Conf.

Results. This section reports the same selective-prediction protocol used in the main text. For each target coverage, we choose a CAS threshold from train out-of-fold predictions and then apply that same numeric cutoff to the held-out test split. The test set therefore does not have to land at exactly the nominal 75%, 50%, or 25% coverage: the score distribution can shift from train to test, and score ties near the cutoff can move a few examples across the boundary. We evaluate Conf at approximately matched test coverage so that the comparison is made at nearly the same operating point. Reporting realized coverage is important because otherwise a small accuracy difference could simply reflect one method answering more examples than the other. Table 9 therefore reports both selective accuracy and realized coverage for CAS and Conf.

A Comparative Analysis of In-Context Learning and Fine-Tuning for Biomedical Information Retrieval and Sentence Extraction Using Research Domain Criteria

Athlene V. Jones
University of North Florida
School of Computing
Jacksonville, FL, USA
N00168025@unf.edu

Khanh Linh Lieu
University of North Florida
School of Computing
Jacksonville, FL, USA
N01583210@unf.edu

Indika Kahanda
University of North Florida
School of Computing
Jacksonville, FL, USA
indika.kahanda@unf.edu

Abstract

Research Domain Criteria (RDoC) is a National Institute of Mental Health framework for studying mental disorders by integrating information across genetics, circuits, and behavior. Manually curating biomedical abstracts relevant to RDoC is a significant challenge due to semantically overlapping construct definitions (e.g., "Acute Threat," "Potential Threat," and "Sustained Threat") and the exponential growth of biomedical literature. This study compares two modeling strategies, domain-adapted fine-tuning and in-context prompting, across two RDoC-related subtasks from the official BioNLP-OST 2019 RDoC shared task. For Task 1, unlabeled PubMed abstracts are retrieved and ranked by relevance to eight of the RDoC constructs. We compare a TF-IDF baseline against ModernBERT and Llama (zero-shot and five-shot) using Mean Average Precision (MAP). For Task 2, the objective is to identify the single most relevant sentence from an abstract for a given construct, evaluated using per-construct accuracy. The fine-tuning track performs end-to-end fine-tuning of BioBERT, PubMedBERT, ModernBERT, and RoBERTa using a cross-encoder input format and per-construct grid search. These are compared against the in-context learning of several open-source language models. Both our approaches are competitive against the best-performing team's score from the BioNLP-OST 2019 RDoC shared task. Taken together, these findings suggest that five-shot prompted LLMs and domain-adapted fine-tuned transformers are viable tools for semi-automating the expert annotation in RDoC curation.

1 Introduction

The exponential growth of biomedical literature poses a significant challenge for mental health and psychiatric researchers and clinicians. PubMed searches for foundational concepts yield results at a scale that makes manual review unmanageable. For example, as of April 2026, the keyword "brain" returns over 2.5 million articles, "depression" returns 724k articles, and "anxiety" yields 397k articles. This challenge is further intensified by the National Institute of Mental Health's adoption of the Research Domain Criteria (RDoC) framework (Insel et al., 2010), which marks a paradigm shift from traditional diagnostic models such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD), to a dimensional, biologically-informed approach to understanding mental disorders. RDoC is comprehensive yet complex, encompassing six major domains of human functioning subdivided into constructs and sub-constructs. Because most of the existing biomedical literature reports (or have been curated) within DSM-ICD terminology, it must be systematically remapped to RDoC concepts, a process that is neither practical nor scalable without computational support.

Effective RDoC-based literature curation requires at least two complementary capabilities. The first is information retrieval (IR): given an RDoC construct, identifying and ranking abstracts from a large corpus that are relevant to it. The second is extracting the most relevant evidence, i.e., sentence extraction (SE): given a relevant abstract, pinpoint the specific sentence(s) that most directly support that construct. These capabilities correspond to the two subtasks established at the BioNLP-OST 2019 shared task (Anani et al., 2019) Task 1 (abstract ranking, evaluated via Mean Average Precision) and Task 2 (sentence extraction, evaluated via per-abstract accuracy), which together define a benchmark for end-to-end RDoC-based literature.

This current study evaluates two computational paradigms across the two aforementioned tasks. For Task 1, the *fine-tuning track* compares pre-trained and fine-tuned BERT family encoders as semantic relevance rankers, while the *in-context learning track* evaluates LLaMA 3.1 (8B) under

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 644–655
July 3-4, 2026 ©2026 Association for Computational Linguistics

zero-shot and few-shot strategies.

For Task 2, the *in-context learning track* applies five small open-source large language models: LLaMA 3.1 (8B), Mistral (7B), Qwen 2.5 (7B), Phi-3 (3.8B), and Gemma 2 (9B) under zero-shot and few-shot strategies with dynamic example selection including based-on TF-IDF, BM25, and sentence embeddings; and the *fine-tuning track* trains four pre-trained transformer encoders: BioBERT, PubMedBERT, RoBERTa, and ModernBERT as per-construct cross-encoder classifiers that jointly encode a construct description and a candidate sentence to produce a binary relevance judgment.

By systematically comparing in-context prompting and supervised fine-tuning, we aim to provide actionable guidance for practitioners seeking to automate RDoC literature curation and to assess the extent to which modern LLMs and fine-tuned transformers can help reduce the expert-annotation bottleneck that has historically constrained progress in this domain (Anani et al., 2019). The RDoC benchmark presents particular challenges not addressed in general-domain NLP settings. The relatively small per-construct training sets and highly overlapping construct semantics make it a demanding low-resource evaluation environment.

This paper makes three main contributions: (1) We present the first systematic comparison of in-context learning and supervised fine-tuning across both RDoC Task 1 (abstract ranking) and Task 2 (sentence extraction), using the original BioNLP-OST 2019 evaluation protocol, (2) We show that few-shot prompting with small open-source LLMs achieves competitive performance with domain-adapted transformers, narrowing the gap to the shared-task best system with no parameter updates, and (3) We provide an empirical analysis demonstrating that lexical example selection (TF-IDF/BM25) consistently competes with dense embeddings for construct-specific prompting in the RDoC setting.

2 Related Work

Biomedical literature curation related to mental health and brain sciences remains a recognized bottleneck in scientific research (Anani et al., 2020). While traditional keyword-matching techniques like TF-IDF (Manning et al., 2008) establish strong initial retrieval baselines, domain-specific extraction requires specialized approaches. To address this, the 2019 BioNLP-OST RDoC Task introduced the inaugural benchmark competition to identify the concepts of Research Domain Criteria (RDoC) in biomedical abstracts (Anani et al., 2019). For abstract ranking (Task 1), Chaudhary et al. (Chaudhary et al., 2019) established the top benchmark (MAP 0.86) using a multi-grain neural relevance ranking system that integrated neural topic models with attention-based sentence interactions. For sentence extraction (Task 2), Laden et al. (2020) developed RDoCer, combining TF-IDF document scoring with a supervised ensemble of lexical and structural features. This approach achieved an accuracy of 0.48, ranking first on the *Sustained Threat*.

In the years that followed, research shifted toward domain-specific transformer architectures, notably with Naseem et al. (Naseem et al., 2022), who demonstrated that BioALBERT could surpass several prior supervised benchmarks by leveraging parameter-efficient pretraining on large biomedical corpora. As the field entered the era of Large Language Models (LLMs), Chen et al. (Chen et al., 2024) evaluated the zero-shot capabilities of models such as GPT-4, finding that although these models exhibit deep clinical reasoning, they often require sophisticated prompting to match the ranking precision of specialized systems. However, neither of these studies incorporates RDoC Task data into the benchmarking process.

To the best of our knowledge, no prior work has directly compared in-context prompting with fine-tuned transformer classifiers on both RDoC Task 1 and Task 2 simultaneously, nor has it evaluated dynamic few-shot example selection strategies in this domain. This study fills that gap with a systematic empirical comparison based on the original shared-task evaluation protocol.

3 Methodology

As depicted in Figures 1 and 2, this study employs two experimental tracks, in-context learning and fine-tuning, applied across the two RDoC task formulations. Both tracks are evaluated against the same benchmark dataset using identical train-test splits and official evaluation protocols, but they represent fundamentally different computational paradigms. The in-context track requires no parameter updates and relies on prompt-based reasoning, while the fine-tuning track adapts pre-trained model weights through supervised training on task-specific labeled data.



Figure 1: This pipeline shows the complete workflow for Task 1. From raw inputs (Title, Abstract, Training set, Query with RDoC Construct and NIMH Definition) through baseline modeling (TF-IDF with Cosine Similarity Ranking), implemented IR techniques (Pre-trained vs Fine-tuned BERT family models), implemented reasoning techniques (Zero-shot vs. Five-shot LLaMA Version 3.1), to final evaluation using Mean Average Precision (MAP).

RDoC Construct	Train	+Test	-Test	Gold
Negative Valence System				
Acute Threat (Fear)	39	53	26	53
Potential Threat (Anxiety)	27	124	15	124
Frustrative Nonreward	21	96	48	96
Sustained Threat	18	82	64	82
Loss	28	90	48	90
Arousal & Regulatory Systems				
Arousal	38	97	11	97
Sleep and Wakefulness	48	121	1	121
Circadian Rhythms	47	123	0	123
Total	**266**	**786**	**213**	**786**

Table 1: Dataset distribution across eight RDoC constructs for training and test splits. Task 1 test abstracts include both positive (relevant) and negative (irrelevant) examples; Task 2 test abstracts are gold-labeled at the sentence level.

3.1 Data

This study employs the benchmark dataset from the BioNLP-OST 2019 RDoC Task (Anani et al., 2019), which consists of human-annotated PubMed abstracts covering eight RDoC constructs. These constructs span two domains of the RDoC framework: the Negative Valence System (Acute Threat, Potential Threat, Sustained Threat, Loss, Frustrative Nonreward) and the Arousal and Regulatory Systems (Arousal, Circadian Rhythms, Sleep and Wakefulness). Gold-standard annotations were produced by three independent expert annotators affiliated with the National Alliance on Mental Illness (NAMI) Montana. (Anani et al., 2019).

The structure of the data set differs between the two tasks, each with a dedicated test set. The shared training set contains 266 abstracts annotated as positive examples for their respective constructs, with per-construct counts ranging from 18 (Sustained Threat) to 48 (Sleep and Wakefulness). The Task 1 test set contains 999 abstracts spanning all eight constructs, of which 786 are labeled as construct-relevant (positive) and 213 as construct-irrelevant (negative). The Task 2 test set contains 244 abstracts, each paired with gold-standard, sentence-level labels that mark the single sentence in each abstract that is most relevant to the target construct. Table 1 reports the full per-construct distribution across all three splits (see Table 1).

For the fine-tuning track's Task 2 experiments, the 266 training abstracts were segmented into individual sentences using NLTK's sentence tokenizer (https://www.nltk.org/), yielding 1,570 sentence-level training instances across all constructs (417 relevant and 1,153 non-relevant). A sentence received a positive label (relevant $= 1$) if it occurred as a substring within the gold relevant context annotation for its abstract; otherwise, it was assigned a negative label (irrelevant $= 0$). This programmatic derivation meant that all sentence-level supervision was obtained directly from the existing abstract-level annotations without additional human labeling. The resulting training data is moderately imbalanced, with approximately 73% non-relevant instances, a characteristic that varies across constructs and bears on training stability.

3.2 In-Context Learning Track

Task 1: Abstract Ranking LLaMA 3.1-70B (AI@Meta, 2024) was queried via the Together AI API (https://www.mintlify.com) to

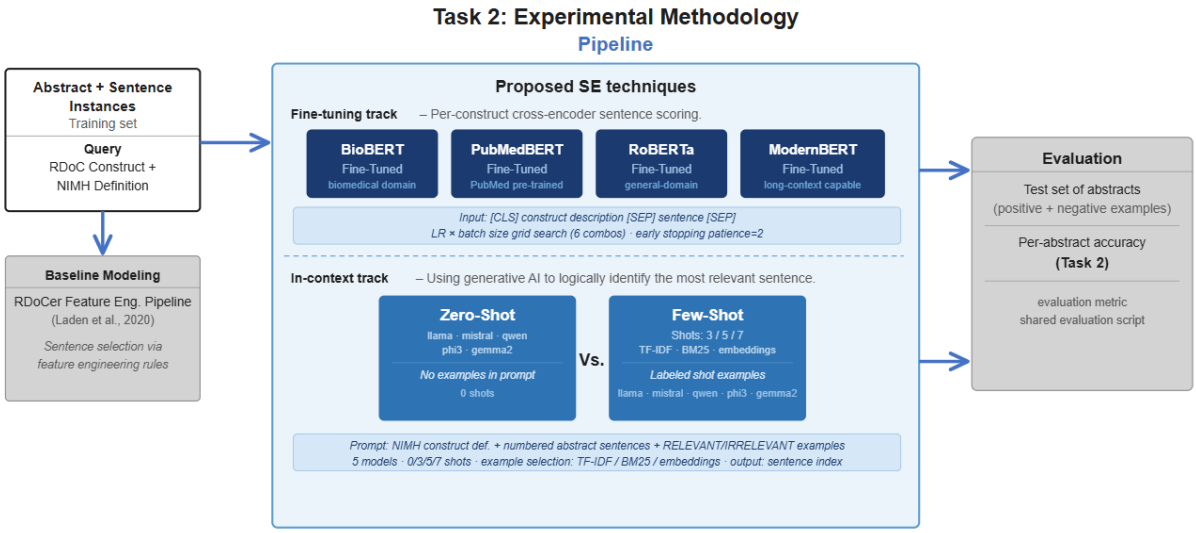


Figure 2: This image depicts the experimental methodology pipeline for Task 2, illustrating two proposed sentence extraction tracks: a fine-tuning track and an in-context learning track comparing zero-shot and few-shot prompting strategies, evaluated using per-abstract accuracy.

assign a relevance score to each test abstract. Two prompting configurations were evaluated. In the zero-shot configuration, the prompt consisted solely of the RDoC construct definition with no labeled examples; five independent trials were conducted, and the results were averaged for robustness. In the five-shot setting, 10 manually selected example abstracts were provided for each construct (five positive and five negative) as in-context demonstrations, enabling the model to discern the linguistic patterns typical of each construct.

Task 2: Sentence Extraction Five small open-source instruction-tuned language models were evaluated locally via Ollama (https://ollama.com/) under zero-shot and few-shot conditions. For each test abstract, the model received a prompt containing the NIMH construct definition and the abstract's sentences enumerated numerically, and was instructed to return a single integer index identifying the most relevant sentence. Under few-shot conditions, labeled RELEVANT and IRRELEVANT sentence examples drawn from the training data were prepended using shot counts of 3, 5, and 7, with demonstrations selected via TF-IDF, BM25, or dense sentence embedding similarity to the test instance. Each configuration was repeated across three random seeds and results were averaged.

3.3 Fine-Tuning Track

Task 1: Abstract Ranking Three BERT-family encoders were evaluated in both pre-trained (frozen) and fine-tuned configurations as abstract relevance rankers, with a binary sequence classification head added to predict whether each abstract is relevant (label $= 1$) or irrelevant (label $= 0$) to a given construct. For fine-tuned variants, a BM25-based pseudo-labeling strategy was first applied to generate construct-specific training subsets. BM25 scores were computed for all training abstracts against an expanded construct query combining the construct name and its NIMH definition. The top-20 abstracts by BM25 score were assigned positive labels; the remaining abstracts received negative labels. This pseudo-labeled subset was used for fine-tuning with early stopping. Pre-trained variants used the same classification head but received no weight updates, serving as a direct comparator.

Task 2: Sentence Extraction. Transformer encoders were fully fine-tuned as per-construct binary classifiers. Rather than encoding the construct and candidate sentence independently, input was formatted using a cross-encoder template that concatenates the NIMH construct definition and the candidate sentence as a single sequence, allowing the model to attend directly across both inputs within a single forward pass. This joint encoding is the core methodological distinction from bi-encoder

or feature-based approaches: the model learns relevance as an interaction between the construct description and the sentence, rather than as a property of either in isolation. A separate classifier was trained for each of the eight RDoC constructs, based on the reasoning that construct-specific fine-tuning is more effective than a single joint model given the semantic distinctiveness of the constructs and the small per-construct training sets.

4 Models

Our approaches were selected to address distinct operational and technical goals within the RDoC annotation pipeline. We chose TF-IDF and BM25 as strong, interpretable lexical baselines to establish a meaningful lower bound for performance. Fine-tuned encoders, including BioBERT, PubMedBERT, ModernBERT, and RoBERTa, were selected to determine if domain-specific pre-training and task-specific supervision could capture subtle semantic distinctions missed by lexical approaches, with the cross-encoder format explicitly conditioning sentence scoring on construct descriptions. Few-shot prompting models evaluated whether instruction-tuned LLMs could generalize RDoC construct identification without task-specific training, representing a low-overhead deployment scenario for resource-constrained clinical settings.

4.1 Baseline

We implement a well-established information retrieval baseline: TF-IDF with Cosine Similarity. For each construct, the query is constructed by concatenating the construct name with its NIMH RDoC construct definition, and all test abstracts are ranked by cosine similarity in the TF-IDF vector space. TF-IDF requires no training and relies solely on keyword overlap between abstracts and construct definitions.

4.2 BM25

BM25 extends TF-IDF with term frequency saturation and document length normalization, producing more stable relevance scores across abstracts that vary substantially in length. In this study, BM25 fulfills two functions: it acts as the retrieval baseline for Task 1 and also provides pseudo-labels used as training supervision for the fine-tuned Task 1 models and for in-context prompting in Task 2.

4.3 In-context Learning Track

4.3.1 Task 1: Abstract Ranking

LLaMA 3.1-70B (AI@Meta, 2024) was queried via the Together AI API (`https://api.together.xyz/`). Its 128K-token context window enables processing of full abstracts alongside construct definitions and in-context examples without truncation. However, its limited exposure to RDoC-specific psychiatric terminology may constrain its ability to distinguish conceptually overlapping constructs, particularly the Threat cluster (Acute Threat, Potential Threat, Sustained Threat).

4.3.2 Task 2: Sentence Extraction

Five open-weight instruction-tuned models were evaluated locally via Ollama: LLaMA 3.1 (8B) (Dubey et al., 2024), Mistral (7B) (Jiang et al., 2023), Qwen 2.5 (7B) (Team, 2024), Phi-3 (3.8B) (Abdin et al., 2024), and Gemma 2 (9B) (Team et al., 2024). These models were selected to represent a range of architectures, parameter scales, and training corpora at the small open-source frontier. LLaMA 3.1 uses grouped query attention with a 128K context window; Mistral employs sliding window attention for memory efficient inference; Qwen 2.5 was pre-trained on over 18 trillion tokens with strong scientific coverage; Phi-3 prioritizes high-quality curated training data over scale; and Gemma 2 uses alternating local and global attention with knowledge distillation from a larger model.

4.4 Fine-tuning Track

4.4.1 Task 1: Abstract Ranking

Three BERT encoders were evaluated in both pre-trained and fine-tuned configurations: ModernBERT (Warner et al., 2024), PubMedBERT (Gu et al., 2021), and BioBERT (Lee et al., 2020). Pre-trained variants compute cosine similarity between construct definition and abstract representations without weight updates. Fine-tuned variants are adapted using BM25 pseudo-labeled training data.

4.4.2 Task 2: Sentence Extraction

The fine-tuning track approached Task 2 as a per-construct binary sentence classification problem. Four pre-trained transformer models were fine-tuned independently: BioBERT (**?**), PubMedBERT (Gu et al., 2021), RoBERTa (Liu et al., 2019), and ModernBERT (Warner et al., 2024), representing biomedical domain-specific, PubMed-specific, general-domain, and

long-context architectures, respectively. All four were fine-tuned using the Hugging Face `AutoModelForSequenceClassification` head with two output labels.

BioBERT and PubMedBERT provide biomedical domain specialization through continued and from-scratch pre-training on PubMed and PMC literature, with BioBERT's retention of the NSP objective offering additional structural alignment with the cross-encoder format. RoBERTa serves as a general-domain baseline, while ModernBERT represents an architecturally modern encoder with an extended 8,192-token context window, trading biomedical depth for up-to-date design choices.

5 Experimental Setup

5.1 Task 1: Abstract Ranking

All models were evaluated on the test set using Mean Average Precision (MAP) as the primary metric. MAP is considered a standard metric for information retrieval tasks, computed as the macro-average across all RDoC constructs. This metric assesses the quality of retrieved abstract rankings, with emphasis on early retrieval of relevant abstracts.

Following the original RDoC Task (Anani et al., 2019), strict separation between training data (Set 1) and test data (Set 2) was maintained, with test data containing positive and negative examples in all constructs examined. 20% of the training data was used as a validation set for parameter tuning. Mean Average Precision (MAP) is computed per construct and then macro-averaged, capturing both precision and ranking quality, which is crucial for real-world curation workflows.

Baseline Comparison: Models were compared against a TF-IDF + cosine similarity baseline, using construct definitions as queries. This baseline represents the state of practice in many curation pipelines. LLaMA inference uses standard sampling to balance consistency and diversity. Both zero-shot and few-shot prompting strategies leverage the model's weights without gradient-based updates. LLaMA was used with default parameters. For few-shot, the optimal number of examples were chosen using a validation set.

For the BERT models, a hyperparameter search algorithm systematically evaluated all combinations of learning rates {1e-5, 2e-5, 3e-5} and epoch counts (2, 3, 4) for each construct independently, resulting in 9 total hyperparameter configurations. Validation performance on construct-specific validation sets guided the selection of the following optimal hyperparameters: Max Sequence Length: 512 tokens, Batch Sizes: 8 (training), 16 (evaluation), Weight Decay: 0.01, Warmup Steps: 50, Learning Rate: Hyperparameter search space (primary: 1e-5 to 3e-5), Number of Epochs: Hyperparameter search space (primary: 2-4 epochs), Early Stopping: EarlyStoppingCallback with patience=2, Optimizer: AdamW.

5.2 Task 2: Sentence Extraction

For the in-context track, five models were evaluated under zero-shot and few-shot conditions. In the zero-shot condition, no labeled examples were included. In the few-shot condition, shot counts of 3, 5, and 7 were evaluated, with examples drawn from the training data using three selection strategies: TF-IDF retrieval, BM25 retrieval, and dense sentence embedding similarity. Each few-shot configuration was repeated on three random seeds, and the results were averaged. The complete experimental grid consisted of 5 models $\times 3$ shot counts $\times 3$ selection strategies $\times 3$ seeds = 135 few-shot runs, plus 5 zero-shot runs.

The prompt structure followed a consistent template across all conditions. The construct definition was provided as a task context block, followed by the abstract sentences enumerated numerically, with the instruction to return a single integer. In the few-shot condition, Example sentences labeled as RELEVANT or IRRELEVANT were inserted with the construct definition and the abstract. (See Figure 5 in the Appendix for a prompt example).

Performance was measured using the official per-abstract accuracy metric via script_task2.py (Anani et al., 2019). A prediction is correct if the selected sentence appears as a substring of the gold Relevant Context annotation. Macro-average accuracy across the 8 constructs is the primary metric.

For the fine-tuning track, Four transformer encoders were fine-tuned per-construct using the Hugging Face Trainer (Wolf et al., 2020) on a Google Colab A100 GPU with mixed-precision (fp16) enabled and a global random seed of 42. For each construct, sentence-level training instances were partitioned 80/20 into training and validation subsets using stratified sampling on the relevance label. Hyperparameters were selected via a grid search over three learning rates $\{1\times10^{-5}, 2\times10^{-5}, 5\times10^{-5}\}$. and two batch sizes (8, 16), yielding six combinations per construct. Early stopping with a patience

of 2 (max epochs of 10) was applied to every grid search run, evaluated once per epoch against validation accuracy. The combination achieving the highest validation accuracy was selected and used to retrain the model on the full training set, again with early stopping.

All runs used AdamW optimization with weight decay of 0.01 and no learning rate warmup. The maximum input sequence length was set to 128 tokens with padding and truncation applied where necessary. During inference, the fine-tuned classifier produced a relevance probability score (the softmax probability of class 1) for each candidate sentence in a test abstract, and the sentence with the highest score was chosen as the prediction. Predictions were evaluated using the official per-abstract accuracy metric via `script_task2.py`, and macro-average accuracy across all eight constructs is the primary reported metric.

6 Results and Discussion

6.1 Task 1: Abstract Ranking

As shown in Figure 3, evaluation across eight distinct approaches revealed substantial differences in performance. The biomedically pre-trained models proved to be competitive but not superior. Across our models, the PubMedBERT models, both the pre-trained and fine-tuned, achieved the highest performance at 0.83 MAP, along with the pre-trained BioBERT. The ModernBERT models performed similarly with MAP scores of 0.82 for the pre-trained and 0.80 for the fine-tuned approach. As for the in-context models, prompting with LLaMA 3.1-70B achieved the lowest performance at 0.79 MAP (zero-shot) and 0.76 MAP (few-shot).

The performance decline from zero-shot to five-shot LLaMA is attributed to selection bias in the manually chosen demonstrations and to lexical overlap within the Threat construct cluster (Acute, Potential, and Sustained Threat), where few-shot examples from one construct inadvertently reinforce vocabulary from the others. In the zero-shot setting, the model relies solely on the NIMH definition, which provides a cleaner disambiguation signal. This suggests that construct-aware, automated example selection is needed for semantically overlapping constructs.

Both our generative reasoning with few examples and discriminative fine-tuning approaches achieve comparable performance on this task. The pre-trained BERT models demonstrated that even without task-specific fine-tuning, the model captures sufficient semantic information for construct classification, where supervised adaptation may not improve results.

Relatively high baseline (TF-IDF) MAP values indicate that, despite high semantic overlap among RDoC constructs, keyword-based retrieval remains competitive. The results demonstrate that while modern general-purpose models are competitive, they do not yet surpass the specialized state-of-the-art for this retrieval task. The RDoC Task's (Anani et al., 2019) best performer (T30) maintains the lead with a MAP of 0.86.

Our models, e.g., ModernBERT, achieve near-perfect results on biologically distinct topics such as Sleep & Wakefulness (0.99) and Potential Threat (0.93), but the performance collapses on overlapping concepts such as Sustained Threat (0.58) and Loss (0.69). This suggests that models effectively capture distinct medical vocabulary but struggle to disentangle nuanced psychological traits such as distinguishing chronic stress from anxiety, dragging down the overall average despite high proficiency in the clearer biological categories.

6.2 Task 2: Sentence Extraction

Neither paradigm, in-context, or fine-tuning reached the RDoCer (Laden et al., 2020), i.e. RDoC Tasks's best (T30) of 0.58, although BioBERT came closest at 0.54. The poor performance of RoBERTa, together with the comparable results of BioBERT and PubMedBERT, suggests that while domain-specific pre-training is required, it alone does not explain performance. Instead, the cross-encoder setup and the use of the construct description as the query appear to be the main methodological factors driving effectiveness.

Fine-tuning The two biomedical domain-specific models produced the strongest results and were nearly identical in performance: BioBERT achieved a macro-average accuracy of 0.543 and PubMedBERT 0.541, both clearing the RDoCer baseline (0.478) by approximately +0.064 and falling within 0.037 of the RDoC Task's best performer (T30) with 0.580. The convergence of these two models supports the conclusion that domain-specific pre-training is the primary differentiating factor, regardless of whether the model is initialized from general BERT weights (BioBERT) or trained from scratch on biomedical text (PubMedBERT). ModernBERT achieved

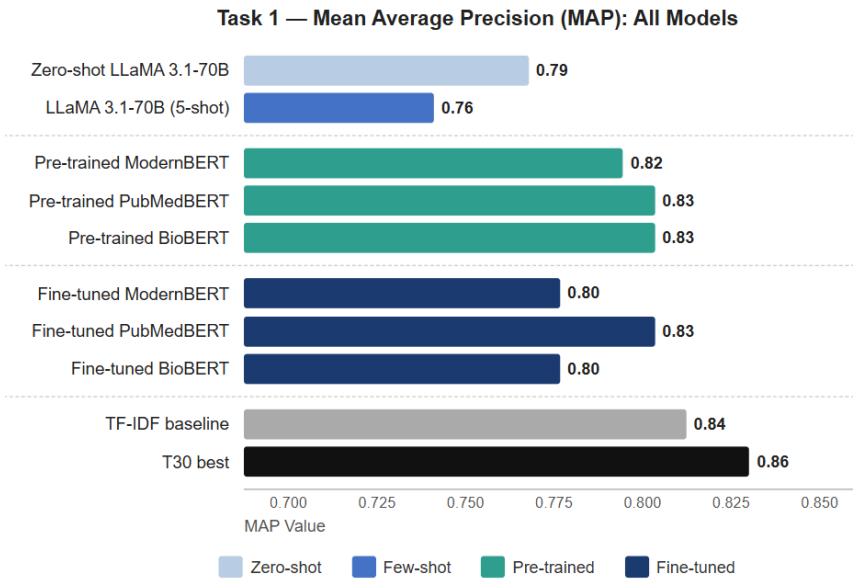


Figure 3: Task 1 MAP results comparing

0.485, marginally clearing the RDoCer baseline (+0.007) but sitting 0.095 below T30. Despite its architectural advantages, its limited biomedical pre-training appears to constrain performance.

RoBERTa was the weakest model at 0.469, falling below both the RDoCer baseline (-0.009) and T30 (-0.111). In four of the eight constructs, the model experienced class collapse: for Arousal, Potential Threat (Anxiety), Acute Threat (Fear), and Frustrative Nonreward, it labeled every test instance as belonging to the non-relevant class. This may be attributable to the removal of the Next Sentence Prediction objective during pre-training, combined with moderately imbalanced training data (73% non-relevant), which together allow the optimizer to converge on a majority-class shortcut. RoBERTa's results should be interpreted with caution, and they are reported for completeness rather than as a reliable performance estimate. Across all models, Acute Threat (Fear) was consistently the hardest construct and Circadian Rhythms the easiest, mirroring patterns from the original 2019 RDoC shared task and consistent with training set size differences across constructs (see Figure 6 in the Appendix) for construct-wise performance.

In-context prompting Zero-shot performance across the five LLMs ranged from 0.380 (LLaMA 3.1 8B) to 0.482 (Gemma 2 9B). All five models fell below the T30 benchmark of 0.580 under zero-shot conditions, and only Gemma 2 exceeded the baseline of 0.40. The spread of more than 10 percentage points between the best and worst zero-shot models suggests that instruction-following capability and model capacity vary meaningfully at this parameter scale for the identification of sentences at the construct-level.

Few-shot prompting produced consistent improvements across all five models, with an average gain of +0.034 over zero-shot. The best overall performance came from Gemma 2 (9B) using 7-shot TF-IDF-based example selection, reaching a macro-average accuracy of 0.517. This is the strongest in-context result and represents a substantial gain over the baseline (+0.117), although it remains 0.063 behind T30. LLaMA 3.1 (8B) showed the largest individual few-shot gain of +0.070, rising from the lowest zero-shot accuracy (0.380) to 0.450 with 3-shot TF-IDF selection, suggesting it is particularly sensitive to prompt context despite weaker zero-shot construct understanding.

Example selection strategy mattered across models. TF-IDF retrieval produced the best few-shot result overall and BM25 was similarly competitive, while sentence embedding selection consistently underperformed both lexical methods. This finding is non-trivial: semantic similarity did not outperform lexical matching for construct-specific example retrieval. We distinguish this finding from prior work on dense retrieval (e.g., MedCPT) (Jin et al., 2023), which targets corpus-level document ranking. Our comparison concerns in-context demonstration selection, where TF-IDF and BM25 excel

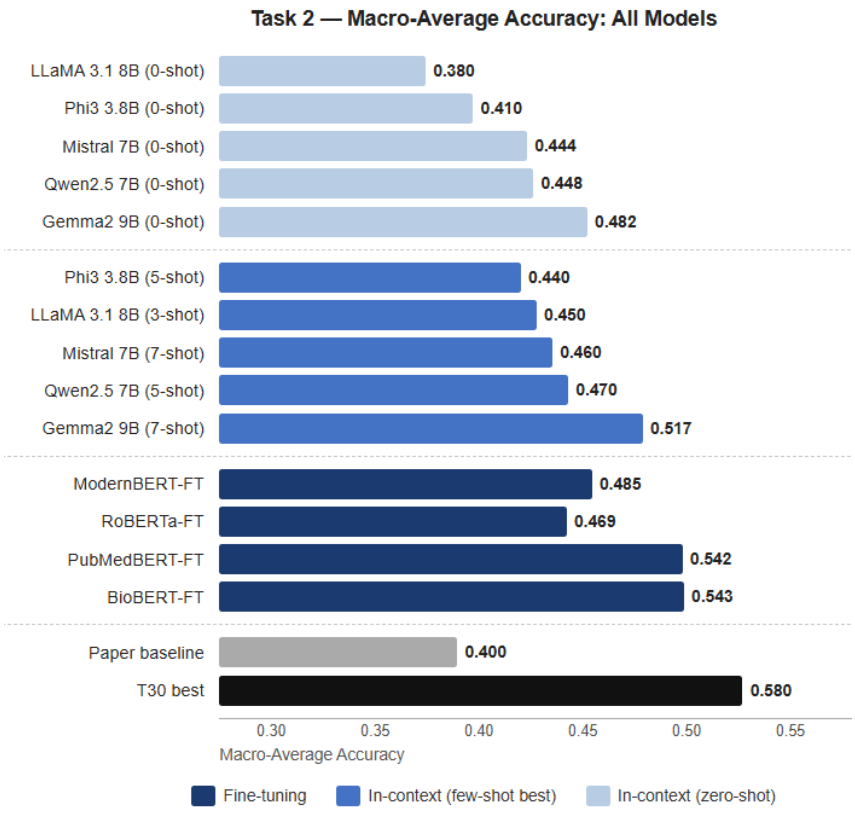


Figure 4: Task 2 Macro-Average Accuracy results

by directly matching National Institute of Mental Health (NIMH) construct-specific terminology through keyword overlap. Conversely, dense embeddings retrieve thematically adjacent examples that often lack this precise, construct-marking vocabulary, ultimately producing less discriminative prompts in this setting.

The consistent underperformance of dense embedding selection (all-MiniLM-L6-v2) across 4 out of 5 models supports this interpretation. For example, TF-IDF retrieves Arousal demonstrations containing terms like 'startle reflex' and 'locomotor activity,' while dense embeddings retrieve sentences from adjacent constructs (e.g., Sleep and Wakefulness) due to overlapping themes, weakening the prompt signal. A similar cross-construct confusion occurs in the Threat cluster, where embedding-selected examples blur the boundary between Sustained Threat and Potential Threat through shared 'chronic stress' vocabulary.

7 Conclusion and Future Work

The results demonstrate that In-context Learning and Fine-tuning offer complementary pathways for RDoC construct-based literature retrieval and extraction, each with distinct strengths. For task 1, PubMedBERT provides marginal performance gains over ModernBERT and BioBERT. LLaMA prompting, despite its relatively low MAP values, offers notable data efficiency, rapid adaptability to new constructs, and reasoning transparency through natural-language explanations, which are critical advantages for applications that require construct-specific justifications. The observed relative performance decline from zero-shot to five-shot LLaMA indicates that construct-aware prompt engineering warrants further investigation.

Future work should focus on construct-specific fine-tuning with more strategic example selection against pretrained models to better capture domain-specific terminology and psychiatric construct language. In addition, conducting an in-depth per-construct performance analysis for both In-context Learning and Fine-tuned models using an ensemble of LLMs would enable a granular comparison of construct-specific strengths and weaknesses across the two approaches.

8 Limitations

This study uses the only known curated dataset related to RDoC via the gold-standard abstracts from the RDoC Task, which was held in 2019. Therefore, our study is restricted by the limited nature of this data. For example, this dataset, is only a single, 6-year-old snapshot of the RDoC Landscape; the RDoC Matrix itself has changed (new constructs like "Sensorimotor Systems" have been added, though no constructs have been removed), not to mention the exponential growth of RDoC-related literature in databases such as PubMed since 2019.

Key challenges for developing methods included the limited training data for certain constructs and the computational cost of hyperparameter tuning with LLMs, which require consistent GPU access and extended training time (that limited our ability use larger models with a very high number of parameters). Task 1 fine-tuning was further constrained by the structure of the available training data, which provides only binary positive or negative abstract assignments per construct rather than ranked relevance judgments. Because this is insufficient to train a ranking model directly, we approximated relevance scores using BM25 as a pseudo-labeling strategy; while this allowed us to apply fine-tuned encoders to the retrieval task, the resulting label noise likely suppresses performance relative to what gold-ranked supervision would yield. We report Task 1 results for completeness but do not claim improvement over the TF-IDF baseline, and we identify the absence of ranked training labels as the primary bottleneck for future work on this task. Another limitation is the lack of construct-specific fine-tuning with more strategic example selection to better capture domain-specific terminology and psychiatric construct language. Multi-task learning was avoided due to severe per-construct imbalance.

In addition, open-source BERT-family models and LLaMA do not represent the full capabilities of the current generation of LLMs (especially their commercial counterparts like GPT), which appear to be improving rapidly. Similarly, the implemented IR techniques do not capture some of the more recent IR innovations like retrieval augmented generation (RAG). In addition, dense retrieval methods such as MedCPT (Jin et al., 2023) were not evaluated, specifically on Task 1, representing a meaningful gap for future work. Lastly, we do not utilize the inherent hierarchical structure of the RDoC Matrix nor the construct-specific matrix elements beyond the construct definitions.

9 Ethical considerations

This work involves automated retrieval of biomedical abstracts related to mental health research using the Research Domain Criteria (RDoC) framework. Our work utilizes publicly available PubMed abstracts and the RDoC shared task dataset, which contains only de-identified scientific literature without patient information or clinical records. No individual-level health data was used in this research. While our system retrieves abstracts related to mental health constructs, we emphasize that the RDoC framework is designed for research purposes to advance understanding of mental disorders through dimensional, cross-cutting approaches. Our work does not diagnose, treat, or make claims about individuals with mental health conditions. We acknowledge the potential for language models to perpetuate stigmatizing language or biases present in the biomedical literature and recommend careful human oversight when deploying such systems in practice.

This study focuses on information retrieval for research purposes and is not intended for direct clinical decision-making. Any future deployment of LLM-based retrieval systems in clinical settings would require rigorous validation, regulatory approval, and appropriate safeguards to ensure patient safety. Clinicians and researchers should critically evaluate retrieved abstracts rather than relying solely on automated rankings. We compare open-source models to promote transparency and reproducibility in biomedical NLP. However, we recognize that LLMs may produce errors, hallucinations, or retrieve irrelevant content. Users of such systems should be trained to identify limitations and maintain appropriate skepticism of automated retrieval results.

Improved retrieval of mental health research literature has the potential to accelerate scientific discovery and improve understanding of mental disorders. However, we acknowledge that advances in automated text mining could also facilitate misuse, such as cherry-picking literature to support predetermined conclusions. We advocate for responsible use of these tools within established evidence-based, scientific and ethical frameworks.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mohammad Anani, Nazmul Kazi, Matthew Kuntz, and Indika Kahanda. 2019. RDoC task at BioNLP-OST 2019. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.

Mohammad Anani, Matt Kuntz, and Indika Kahanda. 2020. Brret: Retrieval of brain research related literature. *AMIA Summits on Translational Science Proceedings*, 2020:53.

Yishai Chaudhary, Pankaj Gupta, and Hinrich Schütze. 2019. Bionlp-ost 2019 rdoc tasks: Multi-grain neural relevance ranking using topics and attention based query-document-sentence interactions. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 227–236. Association for Computational Linguistics.

Qingyu Chen, Jingcheng Du, Yan Hu, and 1 others. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. In *ACM Transactions on Computing for Healthcare (HEALTH)*, volume 3, pages 1–23. ACM.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–751.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Daniel Laden, Shyaman Jayasundara, and Indika Kahanda. 2020. Rdocer: Information retrieval and sentence extraction for mental health using research domain criteria. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 226–229.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In *Bioinformatics*, volume 36, pages 1234–1240. Oxford University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Usman Naseem, Matloob Khushi, Shahadat Uddin Khan, Kamran Shaukat, and Mohammad Ali Moni. 2022. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *Future Generation Computer Systems*, 129:144–153.

Gemma Team, Morgane Riviere, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix: Prompting Example

Figure 5 shows an example of the few-shot prompting format used for both Task 1 and Task 2 experiments.

Sample Prompt Preview

Which sentence provides the strongest evidence that this abstract discusses Arousal?

Definition: Arousal is a continuum of sensitivity of the organism to stimuli, both external and internal. Arousal facilitates interaction with the environment in a context-specific manner (e.g., under conditions of threat, some stimuli must be ignored while sensitivity to and responses to others is enhanced, as exemplified in the startle reflex). It can be evoked by either external/environmental stimuli or internal stimuli (e.g., emotions and cognition). It varies along a continuum that can be quantified in any behavioral state, including wakefulness and low-arousal states including sleep, anesthesia, and coma.

Key indicators: sensitivity to stimuli, startle reflex, behavioral state, locomotor activity, homeostatic drives, arousal level.

Here are examples of **RELEVANT** and **IRRELEVANT** sentences for this construct:

RELEVANT examples:
(1) "valence, arousal), and neuroscience has more recently sought the neurobiological basis of emotions via functional neuroimaging."
(2) "In the present study, we sought to identify intensive valence and arousal affective states via facial EMG activity."
(3) "CONCLUSIONS: This study is the first to identify arousal-specific retrospective affect report discrepancies over time and suggests retrospective reports also reflect personality differences in affective self-knowledge."

IRRELEVANT examples:
(1) "Seventy-seven students ages 19–30 participated in the study."
(2) "Moral standards, as well as conservative beliefs regarding sexuality, are believed to be involved in the etiology and maintenance of this syndrome."
(3) "However, in a recent study (Ehlers et al."

Now identify the relevant sentence(s) in this abstract:

Sentences:
(1) Previous studies have shown that music is a powerful means to convey affective states, but it remains unclear whether and how social context shape the intensity and quality of emotions perceived in music.
(2) Using a with . . .

Total sentences in abstract: 8

Figure 5: Example of the few-shot prompting template used for RDoC construct classification. The prompt includes the task instruction, construct definition with key indicators, example sentence–label pairs (shots), and the target abstract for sentence identification.

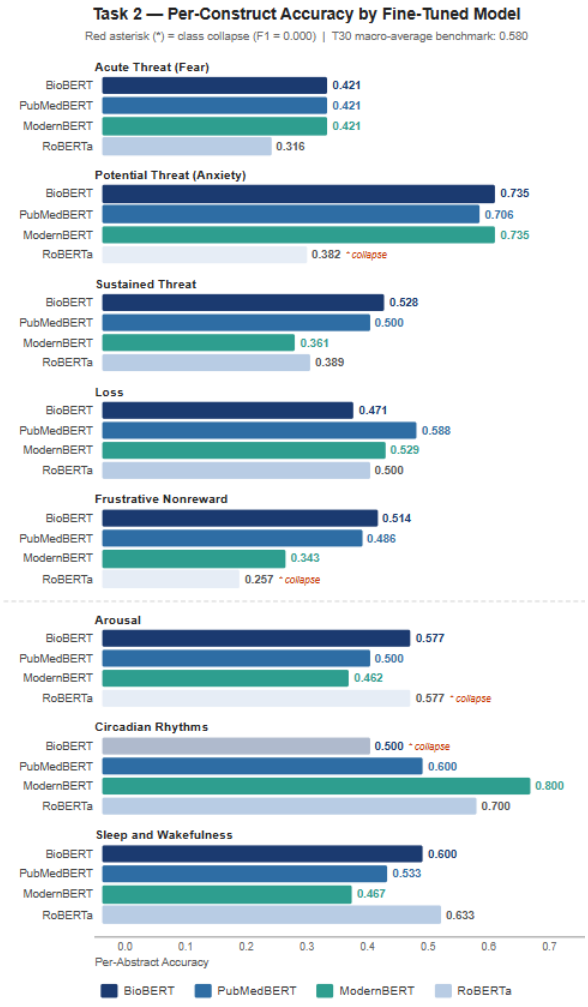


Figure 6: Task 2 Per-construct accuracy for each fine-tuned model across the eight RDoC constructs.

B Appendix: Per-construct Performance

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

Steven Au
Independent Researcher
steventinwing@gmail.com

Abstract

Clinical sources and patient-authored reviews often describe antidepressant side effects in different ways, but these differences are rarely measured directly. We present ClinPeer-AE, a linked dataset for comparing side-effect evidence from PubMed, ClinicalTrials.gov, WebMD, and Drugs.com while preserving source identity. Across five widely prescribed antidepressants, we find low overlap between clinical and peer sources, large differences in relative emphasis, and evidence that many peer-only effects also appear in U.S. Food and Drug Administration Adverse Event Reporting System (FAERS) reports. These findings suggest that patient reviews provide useful context about recurring medication experiences and offer a complementary view of how side effects are described outside formal clinical settings.

1 Introduction

Clinical trials and patient review platforms describe antidepressant treatment in different ways. Trials use controlled designs and standardized adverse-event tables. Reviews show how people talk about treatment in everyday settings, including side effects, discontinuation, and comparisons with other drugs. Both sources matter, but they do not always use the same language.

This paper asks whether those differences can be measured without flattening the sources together. If reviews use different terms, mention effects at different rates, or connect them to specific treatment contexts, that information should remain visible.

We address this with **ClinPeer-AE**[1], a linked dataset that keeps clinical and peer evidence separate while making them comparable. The contribution is not a new adverse-event extractor. Instead, we show how much depends on tracking source, context, and role, especially in reviews where the same phrase may describe an indication, a benefit, a withdrawal symptom, or an on-treatment adverse effect.

Our contributions are:

1. A linked dataset for comparing clinical and patient-authored side-effect evidence for five antidepressants.

2. A set of source-level analyses covering effect coverage, mention salience, condition context, and evidence support.

3. An extraction-quality and sensitivity analysis showing why role classification matters for patient reviews.

2 Related Work

Research on adverse drug effects has been supported by a strong ecosystem of structured resources, benchmark datasets, and biomedical extraction tools. The Side Effect Resource (SIDER) provides curated drug–side-effect associations derived from product labels (Kuhn et al., 2016), while the Medical Dictionary for Regulatory Activities (MedDRA) offers standardized terminology for adverse event reporting in regulatory contexts (Brown et al., 1999). Benchmark corpora such as the Adverse Drug Event (ADE) corpus from medical case reports (Gurulingappa et al., 2012), the CSIRO Adverse Drug Event Corpus (CADEC) from patient forum posts (Karimi et al., 2015), and social media datasets developed for pharmacovigilance (Sarker and Gonzalez, 2017; Feyisetan et al., 2020) have helped define the task in terms of entity recognition, relation extraction, and concept normalization. More broadly, shared tasks in biomedical event extraction have helped shape information extraction methods for clinical and biomedical text (Pyysalo et al., 2012; Kim et al., 2011), and tools such as SciSpaCy have made it easier to build practical pipelines for biomedical parsing and named entity recognition (Neumann et al., 2019).

[1] https://github.com/Prowo/ClinPeer-AE

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 656–664
July 3-4, 2026 ©2026 Association for Computational Linguistics

Adverse-effect evidence, however, is distributed across sources that differ substantially in purpose, language, and reporting style. Clinical trials, regulatory labels, case reports, and patient-authored reviews do not simply present the same information in different formats. They foreground different aspects of medication experience and often vary in granularity, framing, and salience. Clinical trial adverse event tables summarize outcomes observed under controlled conditions, whereas patients often describe effects in more subjective and functionally meaningful terms, such as emotional numbing, cognitive fog, or sexual dysfunction. Prior work has shown meaningful differences between these perspectives, especially for psychiatric medications. The Psychiatric Treatment Adverse Reactions dataset (PsyTAR) provides one of the first dedicated annotated datasets for psychiatric treatment reviews (Zolnoori et al., 2019), and pharmacovigilance studies have shown that social media can surface safety signals that may be underrepresented in formal reporting systems. However, most prior work treats patient-authored text as a supplementary signal for adverse event detection rather than as a source with distinct descriptive and evidential properties.

Recent advances in pretrained and instruction-tuned language models have also improved zero-shot and weakly supervised analysis of medical text. Bidirectional Encoder Representations from Transformers (BERT) family models, including DeBERTa-v3 (Liu et al., 2023) and ModernBERT (Warner et al., 2025), along with instruction-tuned models such as Gemma 4 (Google DeepMind, 2026), make it increasingly feasible to process patient reviews without large task-specific training sets. This is particularly important in patient-authored narratives, where systems must distinguish true adverse effects from related but different phenomena, including treatment indications, perceived benefits, withdrawal symptoms, prior medication history, and background conditions. In this setting, role classification is often necessary to determine how a mention should be interpreted.

3 Data

We assemble a multi-source corpus covering five widely prescribed antidepressants: bupropion, escitalopram, fluoxetine, sertraline, and venlafaxine. The corpus is organized into two source families (Table 1). The *clinical family* comprises PubMed (abstracts and study descriptions) and ClinicalTrials.gov (structured adverse-event tables when available). The *peer family* comprises WebMD and Drugs.com, both of which host patient-authored narratives with condition and rating metadata. All downstream analyses preserve this family label rather than pooling sources before comparison.

Table 1: Corpus composition for the five focal antidepressants.

Source	Docs	Claims	Drugs
PubMed	346	1,871	5
ClinicalTrials.gov	51	29	3
WebMD	18,510	34,670	5
Drugs.com	7,693	39,812	5
Total	**26,600**	**76,382**	–

We use "clinical" to mean formal biomedical evidence broadly, not clinical trials specifically, since PubMed captures a wider range of study designs than trials alone. The released resource is accordingly named **ClinPeer-AE** to reflect this scope. We restrict the corpus to structured review platforms (WebMD, Drugs.com) and biomedical databases (PubMed, ClinicalTrials.gov), excluding social-media sources such as Reddit. Across these four sources, the corpus contains 26,600 documents and 76,382 source-level drug–side-effect claims; downstream metrics aggregate by drug, effect, and source family.

4 Method

4.1 Claim Extraction

Documents are ingested into a shared DuckDB store. A *claim* is an extracted tuple

$$c = (d, e, s, f, r, a, q, x),$$

where d is the drug, e is the effect, s is the source, f is the source family, r is the contextual role, a is the attribution tier, q is confidence, and x is the evidence span. The downstream graph uses the normalized drug–effect pair as the edge key, but retains the source family, source identifier, role, attribution tier, count, and confidence annotations.

The extractor is hybrid rather than a single SciSpaCy model. SciSpaCy provides sentence segmentation, tokenization, biomedical named-entity candidates, and dependency parses. Project-specific

lexicons provide antidepressant aliases and side-effect aliases; these are curated into canonical identifiers before graph construction. Candidate pairs are generated differently by source type. ClinicalTrials.gov adverse-event tables are imported as structured drug–effect rows with available frequency metadata. PubMed abstracts and patient reviews are segmented into sentences; a drug alias and an effect candidate must occur in the same sentence or in a short adjacent-window context, unless the source metadata already fixes the reviewed drug. Dependency distance, trigger terms, negation cues, and lexical match quality determine the confidence score. Each candidate is then assigned an attribution tier encoding how strongly a source frames the effect as medication-related (direct causation > temporal association > co-occurrence > unclear).

To illustrate, consider the sentence "Sertraline caused severe insomnia after two weeks." SciSpaCy segments and tokenizes the input, tags "sertraline" as a drug candidate via the `en_core_sci_md` model, and identifies "insomnia" as a side-effect candidate. The project lexicon maps both to canonical identifiers. The dependency parse links the two entities through the causal trigger "caused," yielding a direct-causation claim for sertraline, insomnia, Drugs.com, peer source, adverse-effect role, confidence 0.92, and the evidence span "caused severe insomnia." By contrast, "Sertraline helped my panic attacks" receives role *benefit* and does not enter the adverse-effect claim set.

For peer sources, we add a review role-classification layer that assigns each candidate drug–effect pair to one of seven context roles: *indication*, *benefit*, *adverse effect*, *withdrawal effect*, *no effect*, *comparator/prior medication*, or *background condition*. Only `adverse_effect` claims pass the role gate in gated configurations. Withdrawal effects are retained as metadata rather than discarded, as discontinuation phenomena may be analytically useful but should not be conflated with on-drug AEs.

Table 2 shows schematic examples of how the role layer changes emitted claims. These examples are paraphrased patterns, not quoted patient reviews.

4.2 Graph Representation

ClinPeer-AE models drugs and side effects as graph nodes, preserving source-family branching at the edge level. Clinical and peer evidence are not merged; the internal graph stores source-specific counts and supports comparative views: clinical-only, peer-only, shared, and high-divergence effects. For each drug, edges to side-effect nodes are annotated with claim counts, average confidence, divergence scores, and rank inversions. An optional FAERS branch adds pharmacovigilance report counts as a third evidence family.

The combined graph for all five drugs contains 87 nodes (5 drug + 82 unique side effects) and 732 branched edges. This design makes source differences measurable without discarding source information.

4.3 Metrics and Hypotheses

We organize the evaluation around four hypothesis tests.

H1: Coverage distinctness. Peer and clinical AE sets have low overlap. We measure this with the Jaccard index (shared / union) and bootstrap 95% CIs. H1 is supported when the upper CI falls below 0.5.

H2: Saliency divergence. Clinical trial AE frequency does not predict peer mention rates. For each shared effect, we compute a *saliency ratio* = peer_mention_count / clinical_trial_frequency_pct. A ratio of 1.0 indicates parity; higher ratios indicate greater relative peer mention salience. We test whether median ratios significantly exceed 1.0 using a Wilcoxon signed-rank test.

H3: Condition-stratified AE profiles. The same drug has different AE mention distributions depending on the treated condition (Depression, Anxiety, OCD, Panic Disorder, PTSD). We test independence using chi-squared contingency tests.

H4: Evidence support by divergence level. Peer-only (high-divergence) effects should carry different evidence backing than shared (low-divergence) effects. We compute a composite support score for each effect combining attribution tier strength (40%), independent peer document support (35%), cross-source peer agreement (25%), reviewer metadata (+0.5), and FAERS corroboration (+0.5). Because most score components are computed over extracted claims, this score is a comparative triage proxy rather than an independent measure of clinical truth. We compare high-divergence ($\geq$0.7) vs. low-divergence ($<$0.3) effects using the Mann–Whitney U test, with Jensen–Shannon divergence and conflict density used as supplementary divergence metrics.

Table 2: Schematic claim extraction examples. The role layer prevents indications, benefits, withdrawal effects, and prior-medication experiences from being counted as ordinary on-treatment adverse effects.

Source pattern	Evidence pattern	Assigned role	Emitted AE claim
ClinicalTrials.gov table	Nausea reported in the sertraline arm with trial frequency	adverse event	(sertraline, nausea)
Patient review	Escitalopram caused insomnia after dose increase	adverse effect	(escitalopram, insomnia)
Patient review	Sertraline helped panic attacks	benefit	none
Patient review	Brain zaps after stopping venlafaxine	withdrawal effect	metadata only
Patient review	Bupropion replaced a prior drug that caused fatigue	comparator/prior medication	none

5 Results

5.1 Extraction Quality

We evaluate extraction on a locked test of 29 Drugs.com reviews and a larger PsyTAR review set, with detailed scores reported in Appendix A. The locked set was produced by one annotator with AI-assisted candidate drafting followed by manual adjudication. Because we did not compute inter-annotator agreement, these scores should be read as a development-quality reliability estimate rather than a final benchmark.

A substantial share of the low F1 comes from role confusion, not only missed effect terms. In the locked test, many false positives are benefit or withdrawal mentions that the extractor counted as adverse effects. This is a central limitation: the downstream comparisons should be read as aggregate analyses over noisy extracted claims, not as verified lists of individual side effects.

We also evaluate separate role-classification baselines. Table 3 shows cross-domain results on PsyTAR review windows, and Appendix B gives additional Gold 120 results. The best local language-model baseline performs much better on adverse-effect role classification than the default extractor, which suggests a clear path for improving the pipeline. It is not yet integrated into the main experiments.

As a stress test, we re-run the peer pipeline with the deterministic role gate enabled. This removes most peer claims, but H1, H2, and H3 remain supported for all five drugs; H4 is more sensitive because its score depends directly on claim volume. This does not solve the extraction problem, but it suggests the main coverage and salience patterns are not driven only by obvious non-adverse review mentions.

Table 3: Cross-domain role classification on 300 PsyTAR overlap windows.

Model	Type	P	R	F1	Acc_7
Gemma 4 E4B	Causal LM	**.893**	.856	**.874**	.573
TF-IDF + LR	Trained	.770	**.966**	.857	**.603**
ClinicalBERT + LR	Trained	.812	.712	.759	.503
BioBERT + LR	Trained	.891	.616	.729	.483
DeBERTa-v3 (NLI)	Zero-shot	.606	.822	.698	.490
TinyLlama 1.1B	Causal LM	.470	.760	.581	.373
ModernBERT (NLI)	Zero-shot	.604	.199	.299	.240
Gemma 4 E2B	Causal LM	.804	.870	.836	.563

Table 4: Coverage distinctness (H1). All drugs show low clinical–peer overlap.

Drug	Clin	Peer	Shared	Jaccard	CI_{95}
Sertraline	102	74	27	.181	[.121, .242]
Fluoxetine	200	68	31	.131	[.089, .177]
Escitalopram	162	72	28	.136	[.092, .184]
Venlafaxine	118	68	24	.148	[.099, .204]
Bupropion	144	67	25	.134	[.086, .188]

5.2 H1: Coverage Distinctness

Table 4 presents coverage results. H1 is supported for all five drugs at the extracted-claim level: Jaccard overlap ranges from 0.13 to 0.18, with upper 95% CIs never exceeding 0.24. At least 76% of the effect space is non-overlapping even accounting for sampling variation. Conversely, 54–65% of peer-reported effects are absent from clinical AE tables.

5.3 H2: Saliency Divergence

Table 5 reports saliency results. H2 is supported for all five drugs (Wilcoxon $p < 10^{-6}$). Patients discuss shared side effects 17–96$\times$ more than clinical trial frequency would predict (Figure 2). Spearman rank correlation is significant for only one drug (fluoxetine), suggesting that saliency ratios capture information not visible in simple rank agreement.

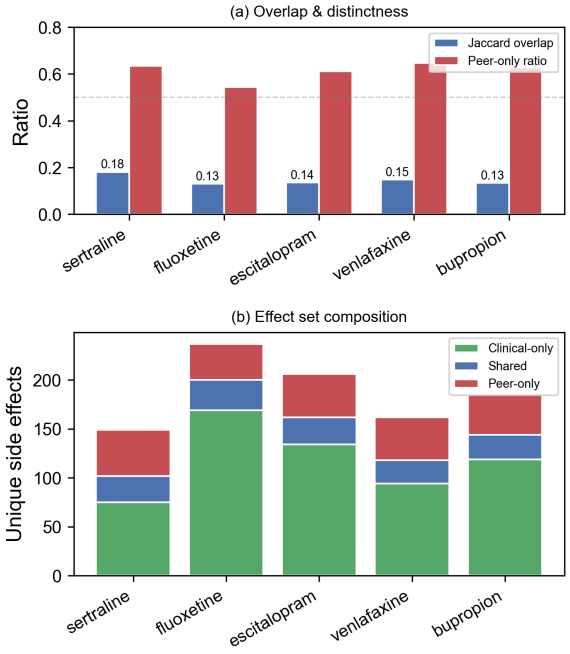


Figure 1: Overlap and effect-set composition by drug. Light regions denote peer-only effects; dark regions denote clinical-only effects.

Table 5: Saliency divergence (H2). Median saliency ratios and Wilcoxon signed-rank tests.

Drug	Shared	Med. ratio	p (Wilcoxon)
Sertraline	28	96.2×	3.7e-09
Fluoxetine	31	41.0×	1.4e-09
Escitalopram	29	56.7×	1.3e-06
Venlafaxine	25	16.6×	6.0e-08
Bupropion	26	68.8×	1.5e-08

5.4 H3: Condition-Stratified AE Profiles

Table 6 presents condition-stratified results. Four drugs show strong evidence of condition-stratified AE profiles; venlafaxine is nominally significant but treated as marginal ($p = 0.046$). Overall, the same drug can yield different AE mention profiles depending on the treated condition.

Notable findings include PTSD patients on venlafaxine reporting weight gain at 24% (vs. 2.9% for depression), bupropion-associated erectile dysfunction concentrated in PTSD reviews, and intrusive thoughts appearing primarily in OCD reviews for sertraline (10.3% vs. $\leq$0.7% for other conditions). Clinical trials typically aggregate AEs across indications; these results suggest that review-derived AE profiles can vary by treatment context. They should not be interpreted as causal estimates of condition-specific drug risk.

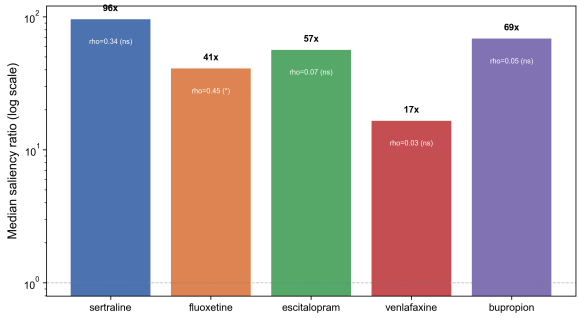


Figure 2: Median saliency ratio by drug (log scale). Dashed line at ratio = 1 (parity).

Table 6: Condition-stratified AE divergence (H3). Chi-squared tests of AE independence across treated conditions.

Drug	p	Top divergent effect
Sertraline	<0.001	intrusive thoughts (OCD 10.3% vs. Anxiety 0.7%)
Fluoxetine	<0.001	panic attacks (Anxiety 13.7% vs. Depression 3.4%)
Escitalopram	<0.001	panic attacks (PTSD 16.2% vs. Depression 4.5%)
Venlafaxine	0.046	weight gain (PTSD 24.0% vs. Depression 2.9%)
Bupropion	<0.001	erectile dysfunction (PTSD 13.5% vs. 0%)

5.5 H4: Evidence Support by Divergence Level

Table 7 presents H4 results. The hypothesis is supported for 4/5 drugs, but in the *opposite* direction: peer-only (high-divergence) effects receive lower support scores than shared (low-divergence) effects. The delta is consistently negative (-0.14 to -0.38). Because the score draws on extracted claim volume and pipeline-derived attribution, this result is best read as a suggestive comparative pattern rather than independent validation of evidential strength.

This finding is useful mainly for triage. It suggests three things: (1) divergence correlates with weaker measured support in this graph; (2) not all peer-only effects are equal, since some (e.g., *panic attack*, *brain zaps*) have higher support because they appear across platforms and in FAERS, while others may reflect extraction or reporting noise; (3) support scoring can prioritize which peer-only effects merit manual review and follow-up analysis.

FAERS triangulation. FAERS data (2021–2025, 36,826 cases) shows that 42–56% of peer-only effects also appear in suspect-drug pharmacovigilance reports, with the highest rate for venlafaxine (55.9%) and the lowest for fluoxetine (42.5%). This is external triangulation, not causal validation: FAERS is spontaneous-report data and can share reporting biases with review platforms. Still, the overlap suggests that many peer-only effects are

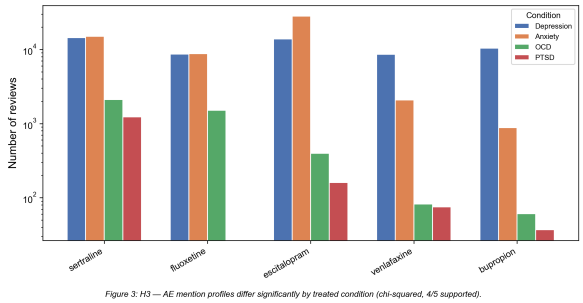


Figure 3: Condition-stratified AE review counts by drug.

Table 7: Evidence support by divergence level (H4). Peer-only effects receive consistently lower support scores than shared effects.

Drug	Peer-only	Shared	Δ	p
Sertraline	52	21	−0.27	0.0008
Fluoxetine	40	28	−0.38	4.9e-06
Escitalopram	60	12	−0.36	0.0003
Venlafaxine	59	9	−0.14	0.150
Bupropion	52	15	−0.22	0.011

not isolated review artifacts, while the support gap shows they remain thinner than shared effects on average.

Overall, most hypothesis tests reach nominal statistical significance across the five drugs. The main exception is venlafaxine under H4 ($p = 0.15$), where the direction is consistent but the sample of shared effects is small ($N = 9$). The venlafaxine H3 result is significant but marginal and should not be overinterpreted.

6 Discussion

The results suggest that differences between formal clinical evidence and patient reviews are not entirely random. Across five antidepressants, peer sources provide broader side-effect coverage and include recurring experiential terms such as brain zaps, emotional blunting, and intrusive thoughts that are weakly represented in clinical trial AE tables.

The H4 result runs counter to our initial expectation. We expected peer-only effects with high divergence to have stronger converging evidence, such as multiple independent reviews, cross-platform support, or FAERS corroboration. Instead, shared effects receive higher support scores. This pattern is suggestive but not definitive, since the score itself depends on extracted claim volume. It mainly shows that effects documented by both clinical and peer sources tend to accumulate more evidence in this graph. Peer-only effects often also appear in FAERS (42–56%), but they are thinner in this evidence base on average, making the score useful for triage rather than clinical adjudication.

Implications for extraction. The findings point to a practical need in patient-review analysis: extraction systems should model role and source explicitly. A simple drug–effect label cannot tell whether a phrase describes an indication, a benefit, a withdrawal symptom, a current side effect, or an experience with a prior medication. Evaluation should also keep clinical and review language separate enough to see where systems fail. ClinPeer-AE provides a testbed for that kind of role-aware extraction and source-level comparison.

One practical implication is that early source pooling hides the differences this paper is trying to measure. Keeping the sources separate makes those differences visible.

Evidence asymmetry. The peer corpus is substantially larger than the clinical-trial portion, and some observed divergence reflects study-design and reporting differences rather than purely biological difference. Clinical trials use strict inclusion criteria, controlled dosing, and structured AE capture; patient reviews reflect heterogeneous populations, variable adherence, and unstructured narrative. We discuss this as a limitation below, not as a flaw; the framework is designed to measure observable source differences, not to adjudicate biological truth.

7 Conclusion

ClinPeer-AE shows that clinical evidence and patient-authored reviews can be compared without merging them into one evidence pool. Across five antidepressants, the two source families differ in coverage, salience, condition context, and measured support. The main value of the dataset is therefore not a definitive list of side effects, but a reproducible way to study how adverse-event language and evidence vary across reporting settings.

Future work should improve extraction quality, add multi-annotator evaluation, expand clinical source coverage, and integrate stronger role classification before using peer-only effects as pharmacovigilance signals.

8 Limitations

Extraction F1 ranges from 0.11 to 0.21, primarily due to role confusion, including benefit and withdrawal mentions misclassified as AEs. The comparative framework does not eliminate this problem because source families have different language and therefore different error profiles. The Gemma 4 role classifier offers a clear improvement path, but is not yet integrated into the default pipeline.

The locked evaluation set was produced by a single annotator with AI-assisted candidate drafting and manual adjudication. No inter-annotator agreement was computed. The extraction scores therefore quantify performance against a useful but limited gold set rather than a mature shared-task benchmark.

We compare the production extractor to a lexicon baseline and a deterministic role gate, but we do not benchmark against strong supervised or generative extraction systems. A single extraction pipeline can also favor one source family over another, since clinical tables, PubMed abstracts, and patient reviews use different language. The NLP component should therefore be read as an enabling pipeline for source-aware analysis, not as a competitive extraction contribution.

WebMD and Drugs.com contain far more documents than ClinicalTrials.gov. Volume normalization helps but does not eliminate the asymmetry, and future work should incorporate additional clinical sources.

Patient reviews capture association, not causation, and FAERS reports reflect spontaneous reporting rather than controlled observation. The framework measures observable source differences; it does not establish which source is correct.

Patient reviews come from a self-selected population that may overrepresent negative experiences. The exclusion of Reddit also limits generalizability to broader peer discourse.

Five drugs across four hypotheses yield 20 tests without family-wise correction. Individual p-values should therefore be interpreted cautiously. Chi-squared tests for rare conditions such as OCD and PTSD also involve small cell sizes, and the venlafaxine H3 result ($p = 0.046$) should be treated as marginal.

9 Ethics Statement

This work analyzes publicly available evidence sources (PubMed, ClinicalTrials.gov, WebMD, Drugs.com, and FAERS) without identity linkage across platforms. Patient-review evidence is treated as observational and self-reported, not as causal ground truth. The goal is to quantify source-specific differences between formal evidence and patient-authored reviews in an auditable, reproducible manner, not to replace clinical evidence or to make treatment recommendations. All data is used in aggregate; no individual reviews are quoted or identifiable.

References

Elaine G. Brown, L. Wood, and S. Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Safety*, 20(2):109–117.

Oluwaseun Feyisetan, Tom Diethe, and Suzanne Kasarda. 2020. Privacy-preserving word representations for improving pharmaceutical post-market surveillance. In *Proceedings of the 3rd Clinical NLP Workshop*, pages 79–85.

Google DeepMind. 2026. Gemma 4. Technical Report.

Harsha Gurulingappa, Anand M. Rajput, Alasdair Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support information extraction for adverse drug effects. In *Proceedings of BioNLP*, pages 73–80.

Sarvnaz Karimi, Alvaro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA. Association for Computational Linguistics.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079.

Pengfei Liu, Yuan Cao, Min Lin, Yijun Li, and Xudong Liang. 2023. DeBERTa-v3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. SciSpaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Overview of the ID, EPI and REL tasks of BioNLP shared task 2011. *BMC Bioinformatics*, 13(Suppl 11):S2.

Abeed Sarker and Graciela Gonzalez. 2017. A corpus for mining drug effects from social media. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 282–289.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The psytar dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.

A Extraction Performance

This section gives the numeric extraction scores behind Section 5.1. The table is included as diagnostic context because the main analyses depend on extracted claims, but the detailed extractor comparison is not itself the central result.

Table 8: Extraction performance by configuration.

Config	Set	P	R	F1
scispacy_hybrid	Locked (29)	.174	.268	.211
scispacy_hybrid	PsyTAR (647)	.259	.071	.111
baseline lexicon	PsyTAR (647)	.256	.072	.112
scispacy + rules	PsyTAR (647)	.348	.005	.009

B Role Classification Baselines

Table 9 reports Gold 120 results for the seven-role classification task (locked_test + hard_case_dev splits). The PsyTAR cross-domain comparison is reported in the main paper in Table 3. "Trained" classifiers use PsyTAR overlap data; "Zero-shot" models require no task-specific training.

C Cross-Drug Peer-Only Effects

This section lists the recurring peer-only effects discussed briefly in the main results. It is kept in the appendix so the main paper can focus on the aggregate tests rather than a long effect inventory.

Nineteen effects are peer-only across all five drugs: *panic attack*, *brain zaps*, *emotional blunting*, *discontinuation syndrome*, *depersonalization*, *erectile dysfunction*, *excessive yawning*, *intrusive thoughts*, *anxiety attack*, *jaw clenching*, *numbness*, *back pain*, *muscle spasms*, *shaking*, *brain fog*, *delayed ejaculation*, *night sweats*, *confusion*, and *stomach pain*. An additional 13 effects appear in 4/5 drugs, including *tinnitus* and *agitation*. At the aggregate level, these terms recur across peer sources; however, individual effects still require manual and clinical validation before being treated as safety signals.

D Support Score Components

The composite support score is computed as:

- **Attribution tier** (40%): direct causation (1.0) $>$ temporal association (0.75) $>$ co-occurrence (0.50) $>$ unclear (0.25).
- **Independent peer documents** (35%): $\log(1+$ distinct_review_count$) \times 10$.
- **Cross-source support** (25%): binary flag for presence in both WebMD and Drugs.com.
- **Metadata support** (+0.5): review includes duration, condition, or demographics.
- **FAERS support** (+0.5): drug–effect pair appears in FAERS suspect-drug reports.

Total score range: 0 to 2.0.

E Supplemental: Attribution Language Comparison

The original H4 formulation compared attribution language between clinical and peer sources. All five drugs showed significant differences ($p < 10^{-36}$), but the effect sizes were small (Cliff's δ 0.07–0.20) and largely reflected genre differences: clinical trial AE tables list effects statistically ("treatment-emergent adverse event") while patient reviews use causal framing ("this drug caused nausea"). This result is technically significant but substantively uninformative and is reported here as supplemental.

Table 9: Review role classification on Gold 120 (Drugs.com reviews, 7 roles). Adverse-effect binary P/R/F1 and 7-class accuracy.

Model	**Type**	**Split**	**N**	**Adv P**	**Adv R**	**Adv F1**	**Acc$_7$**
Rules v1	Deterministic	pilot	148	1.00	.375	.545	.433
Gemma 4 E4B	Causal LM	locked_test	261	.889	.857	.873	.628
Gemma 4 E4B	Causal LM	hard_case_dev	367	.920	.896	.907	.676
Gemma 4 E2B	Causal LM	locked_test	261	.691	.839	.758	.628
Gemma 4 E2B	Causal LM	hard_case_dev	367	.803	.852	.827	.681
TF-IDF + LR	Trained	locked_test	261	.533	.429	.475	.448
TF-IDF + LR	Trained	hard_case_dev	367	.625	.652	.638	.425
ClinicalBERT + LR	Trained	locked_test	261	.326	.268	.294	.310
ClinicalBERT + LR	Trained	hard_case_dev	367	.637	.443	.523	.395
BioBERT + LR	Trained	locked_test	261	.325	.232	.271	.276
BioBERT + LR	Trained	hard_case_dev	367	.657	.565	.607	.373
DeBERTa-v3 (NLI)	Zero-shot	locked_test	261	.341	.821	.482	.391
DeBERTa-v3 (NLI)	Zero-shot	hard_case_dev	367	.512	.730	.602	.463
ModernBERT (NLI)	Zero-shot	locked_test	261	.377	.929	.536	.341
ModernBERT (NLI)	Zero-shot	hard_case_dev	367	.478	.748	.583	.436
TinyLlama 1.1B	Causal LM	locked_test	261	.189	.696	.298	.180
TinyLlama 1.1B	Causal LM	hard_case_dev	367	.308	.809	.446	.267

A Deterministic Multi-Stage Retrieval Pipeline for Longitudinal EHR Question Answering

Shubham Agarwal[1], Thomas Searle[1], Ninoslav Majkic[2], Niko Muller-Grell[1], Richard Dobson[1,3]

[1]King's College London, London, UK
[2]South London and Maudsley NHS Foundation Trust, London, UK
[3]Health Data Research UK and University College London, London, UK
Correspondence: shubham.agarwal@kcl.ac.uk

Abstract

Retrieval-augmented generation (RAG) holds promise for clinical question answering over electronic health records (EHRs), but existing systems treat retrieval as an opaque subroutine, limiting auditability and reliability in patient care workflows. We introduce a deterministic multi-stage retrieval pipeline for longitudinal EHR question answering that decomposes retrieval into four distinct, ablated stages where each stage is instrumented with diagnostic metrics, making the flow of clinical evidence measurable and auditable at every step. Evaluated on a broad LLM-annotated cohort and an expert-annotated cardiovascular benchmark developed alongside clinicians from real ICU records, the full pipeline achieves 22-23% relative recall gain over a strong dense retrieval baseline across both cohorts, with consistent improvements in downstream answer quality. The pipeline's deterministic and transparent design addresses a critical gap in clinical NLP: retrieval systems that clinicians and researchers can not only rely on, but inspect, audit, and build upon for real-world deployment.

1 Introduction

Electronic Health Records (EHRs) capture rich longitudinal information about patients, including demographics, diagnoses, medications, and procedures, but a large fraction of this information exists only in unstructured free text: discharge summaries, nursing reports, radiology narratives, and echo reports (NHS, 2023; Häyrinen et al., 2008). Unlike coded fields, these narratives capture nuanced clinical reasoning, temporal relationships, and contextual detail that structured data systematically omits (Menachemi and Collum, 2011). A single patient admitted to an intensive care unit can accumulate hundreds of notes across multiple encounters, and over a longitudinal care history this figure reaches into the thousands (Zhang et al., 2017). The result is that clinically decisive information spanning areas such as trial eligibility (Raghavan et al., 2014), diagnosis (Shivade et al., 2014), and medication safety is present in the record but practically inaccessible at scale.

Conventional query interfaces and keyword search fail in this setting because they ignore clinical context, cannot resolve synonyms and abbreviations, and do not scale to long, redundant, and noisy documentation (Hill et al., 2021; Spasic et al., 2020). What clinicians and researchers need instead are systems that can answer natural language questions directly over a patient's record: whether a patient has ever met criteria for a particular syndrome (Pathak et al., 2013), how a clinical state has evolved across admissions (Silva and Matos, 2022), or why a medication was started, adjusted, or discontinued (Li et al., 2011). The core challenge is not data storage or extraction but reliable, traceable retrieval: identifying the specific note segments that support an answer, across a large and heterogeneous longitudinal record, without missing evidence that could affect patient care.

Retrieval-augmented generation (RAG) has emerged as a promising framework for this challenge, combining a retrieval component with a generative model conditioned on retrieved context (Lewis et al., 2020). In the EHR setting, RAG systems retrieve relevant passages from a patient's longitudinal notes and generate concise, grounded responses to clinical queries (Sivarajkumar et al., 2024; Alkhalaf et al., 2024), making the approach well suited to long, noisy, and heterogeneous documentation. However, the quality of generated answers depends directly on what is retrieved, as a generator conditioned on incomplete evidence will produce answers that are fluent but miss critical clinical information. Despite this, the question of how to engineer retrieval for high recall and auditability under realistic clinical constraints remains largely open.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 665–678
July 3-4, 2026 ©2026 Association for Computational Linguistics

In this work, we address this gap with a deterministic multi-stage retrieval pipeline for longitudinal EHR question answering, designed around the constraints of real clinical workflows. The pipeline was developed and evaluated alongside clinical experts, with an expert-annotated cohort constructed specifically for cardiovascular medication review, a high-stakes setting where incomplete evidence recovery can directly compromise clinical decisions. Specifically, we:

- *Decompose retrieval into four distinct, ablated stages* that narrow the clinical corpus, expand query coverage, recover missed evidence, and optimise final ranking to ensure no clinically relevant information is lost.
- *Instrument each stage with diagnostic metrics*, ensuring the flow of clinical evidence is measurable and auditable at every step, enabling systematic improvement of the pipeline and, ultimately, reliable clinical use.
- *Validate on an expert-annotated cardiovascular cohort developed alongside clinicians from real ICU records and an LLM-annotated cohort*, achieving over 22% relative recall gain over a strong dense retrieval baseline across both cohorts with consistent improvements in downstream answer quality.

2 Related Work

Clinical question answering over EHRs requires systems to identify small sets of highly relevant, patient-specific evidence from large, heterogeneous, and longitudinal records. Unlike open-domain Question-Answering (QA), clinical settings impose strict requirements on evidence traceability, recall of critical findings, and bounded result sets suitable for clinician review since the retrieved evidence may directly affect patient care and must therefore be reliable and auditable (Roberts et al., 2017; Koopman et al., 2015). In this context, system reliability therefore depends not on answer fluency alone, but on whether the correct note segments are retrieved in the first place which is a challenge that has increasingly motivated retrieval-augmented approaches in clinical NLP.

Large Language Model (LLM) systems for clinical tasks frequently adopt RAG to ground outputs in external evidence (Lewis et al., 2020; Izacard and Grave, 2021; Guu et al., 2020). In medicine, such approaches have been applied to clinical question answering (Singhal et al., 2023; Lehman et al., 2023), summarization (Nazi and Peng, 2024), and decision support (Jin et al., 2021). However, evaluation in these systems typically emphasizes end-to-end answer correctness or textual quality, while the retrieval component itself receives little direct analysis.

Deterministic Information Retrieval (IR) approaches for EHRs often leverage medical ontologies such as SNOMED CT (Donnelly, 2006) and Unified Medical Language System (UMLS) (Bodenreider, 2004) to improve recall and interpretability (Sivarajkumar et al., 2024) . Named entity recognition and concept normalization systems such as cTAKES (Savova et al., 2010), MetaMap (Aronson, 2010), and MedCAT (Kraljevic et al., 2021a) have been widely used to attach structured semantics to clinical narratives. Ranking models, which score and order candidate passages by relevance to a query, range from traditional probabilistic methods such as BM25 (Robertson and Zaragoza, 2009) to neural dense retrievers adapted to biomedical text (Karpukhin et al., 2020; Lee et al., 2020; Gu et al., 2021). Despite these advances, most clinical IR pipelines remain relatively shallow, typically involving one or two retrieval stages followed by ranking, and sparse methods such as BM25 can struggle in clinical settings where the same concept can be expressed through varied terminology, abbreviations, and paraphrasing, motivating the shift towards dense retrieval methods and more flexible, LLM-driven retrieval frameworks.

Modern RAG systems increasingly embed retrieval inside LLM-driven control loops. In these architectures, LLMs may reformulate queries (Asai et al., 2024), generate intermediate hypotheses (Yao et al., 2023), judge evidence relevance (Asai et al., 2024), select passages (Yao et al., 2023), or determine when to stop retrieving (Jiang et al., 2023; Fan et al., 2024; Gao et al., 2023b). Variants of iterative (Shao et al., 2023; Trivedi et al., 2023), recursive (Kim et al., 2023) and agentic (Shinn et al., 2023) retrieval have been proposed to improve reasoning over multi-hop or ambiguous queries.

While LLM-driven retrieval introduces flexibility through model-guided query reformulation, passage selection, and iterative reasoning, it shifts control and evidence selection to opaque model-internal decisions, reducing auditability and in-

troducing variable latency and computational cost that complicate integration into time-sensitive and privacy-sensitive clinical workflows. Taken together, existing work presents two extremes: deterministic but relatively shallow clinical IR pipelines, and flexible but opaque LLM-orchestrated retrieval frameworks. What remains missing is a systematic treatment of retrieval itself as a multi-stage, diagnosable process for patient-level QA.

3 Methodology

3.1 Dataset

This study uses the MIMIC-III (Johnson et al., 2016) clinical database, a publicly available collection of deidentified electronic health records from intensive care units at the Beth Israel Deaconess Medical Center. MIMIC-III contains unstructured clinical documentation such as admission notes, progress notes, nursing notes, discharge summaries, radiology reports, and other narrative reports associated with ICU and related hospital encounters.

For this work, unstructured free-text notes from the NOTEEVENTS table were used, yielding longitudinal collections that reflect realistic documentation patterns across multiple admissions and care settings.

3.2 Problem Formulation

Each patient record consists of multiple hospital visits, and each visit comprises a collection of free-text clinical notes.
For a given patient i, the record can be written as:

$$R_i = \{V_{i1}, V_{i2}, \ldots, V_{iM_i}\},$$

where V_{im} denotes the m-th visit for patient i, and M_i is the number of visits. Each visit V_{im} is itself a set of notes

$$V_{im} = \{n_{im1}, n_{im2}, \ldots, n_{imK_{im}}\},$$

where n_{imk} is an unstructured free text clinical note associated with that visit, and K_{im} is the number of notes for visit m of patient i. Each note n_{imk} is associated with a clinical category c, where

$$c = \{\text{discharge, nursing, ECG, radiology, echo}\}.$$

The objective of this work is to enable patient-centered QA over these note collections. Given a natural language query q about a specific patient i, for a specific visit m, the system must produce a natural language answer a, and a set of supporting evidence segments E_i drawn from notes from one visit across one or multiple categories. In the RAG setting, a retriever g selects relevant note chunks and a generator h produces the final answer conditioned on them.

In our experiments, queries are evaluated under two complementary settings: a category-restricted setting, where E_i is limited to notes of a specific clinical category c, and a visit-level setting, where E_i spans all note types within a visit. Both settings reflect realistic clinical workflows and are evaluated throughout this work.

Notation note For clarity, we present the formalism for the category restricted setting within a single patient visit (i, m, c). The visit-level setting arises naturally by expanding the candidate corpus to $\bigcup_c \mathcal{C}_{imc}$. The proposed pipeline is not restricted to this scope, and cross-visit retrieval can be performed by expanding the candidate corpus to $\bigcup_m \mathcal{C}_{imc}$, without modifying the underlying retrieval or representation components.

3.3 Modules

3.3.1 Chunking

Clinical notes vary substantially in length and structure, and naïvely splitting them into fixed-size segments risks either breaking coherent clinical units or creating overly long passages. We adopt a structure-aware chunking strategy that respects natural note boundaries and category-level organisation.
For a given patient i and visit m, the visit-level note collection is $V_{im} = \{n_{im1}, \ldots, n_{imK_{im}}\}$, where each n_{imk} is an unstructured clinical note with an associated category c. Chunking operates at the level of a patient–visit–category triplet in all experimental settings as notes from different clinical categories are chunked independently since merging across categories would conflate clinically distinct documentation types and degrade retrieval signal. For a fixed (i, m, c), $\mathcal{N}_{imc} = \{n_{imc1}, \ldots, n_{imcK_{imc}}\}$ denote the sequence of notes of category c. From this sequence, we construct a sequence of chunks
$\mathcal{C}_{imc} = \{u_{imc1}, \ldots, u_{imcT_{imc}}\}$, where each chunk u_{imct} is a concatenation or sub-span of one or more elements of $\mathcal{N}_{imc}$ subject to a character-length constraint $L_{min} \leq |u_{imct}| \leq L_{max}$, with $L_{min} = 900$ and $L_{max} = 1500$.

This strategy, as outlined in Appendix A.1, outperforms purely length-based recursive segmen-

tation, as confirmed by ablations reported in Appendix A.2 (Table 3), and yields retrieval units that respect natural note boundaries, are easier to interpret clinically, and remain short enough to be handled efficiently.

3.3.2 Metadata Enhancement via NER

For each note chunk u_{imct}, we enrich the raw text with structured metadata derived from clinical Named Entity Recognition (NER). We apply the MedCAT (Kraljevic et al., 2021b) model to identify clinical entities (e.g., disorders, findings, medications) and leverage its built-in SNOMED CT (SNOMED) mapping to assign each detected entity e to an ontology concept with corresponding semantic category $s(e) \in \mathcal{S}$, where $\mathcal{S}$ denotes the space of clinical semantic types.

Formally, the NER model produces a set of entity annotations $\mathcal{E}_{imct} = \{(e_j, s(e_j))\}_{j=1}^{J_{imct}}$, where e_j is the entity span and $s(e_j)$ its semantic type. We then compute summary metadata for the chunk as

$$M_{imct} = \Big(\big(|\{e_j \in \mathcal{E}_{imct} : s(e_j) = \tau\}| \big)_{\tau \in \mathcal{S}}, \mathcal{E}_{imct} \Big)$$

which stores both entity counts per semantic type and the full list of detected entities. These metadata M_{imct} are stored alongside the chunk text u_{imct} in the retrieval index.

Attaching ontology-based categories to each chunk enables the pipeline to rapidly narrow a large clinical corpus to a semantically plausible subset before retrieval stages. For example, a query about medications can be directed towards chunks containing entities labeled as clinical drugs or medicinal substances, reducing noise and computational cost for downstream stages.

3.3.3 Embedding and Indexing

Each chunk, augmented with its MedCAT-derived metadata is encoded using a dense embedding model to produce a fixed-dimensional vector representation suitable for semantic similarity search (Karpukhin et al., 2020; Ni et al., 2022; Izacard and Grave, 2021). These embeddings are computed for all chunks in a patient's corpus and stored in a vector database that supports efficient approximate nearest-neighbor search. At query time, this index enables rapid retrieval of the most semantically similar chunks, serving as the foundation for the subsequent stages in the pipeline.

We use the `Chroma`[1] vector database and the `BAAI/llm-embedder` (Zhang et al., 2023) model as the encoder for both queries and chunks, as it has shown strong performance on semantic retrieval benchmarks.

[1] https://www.trychroma.com

3.3.4 Query Category Generation

To guide metadata filtering of the corpus, we generate relevant SNOMED CT semantic types from each query q. Using MedGemma 4B (Sellergren et al., 2025), a medically pre-trained language model, with few-shot prompting over $\mathcal{S}$, a set of candidate ontology categories curated from the SNOMED CT hierarchy in consultation with clinical experts, we produce two disjoint sets of semantic types: $\mathcal{S}_q^d \subseteq \mathcal{S}, \quad \mathcal{S}_q^c \subseteq \mathcal{S}$, where $\mathcal{S}_q^d$ denotes direct categories explicitly referenced in q (e.g., 'clinical drug', 'medicinal product' for medication queries), and $\mathcal{S}_q^c$ denotes complementary categories that frequently co-occur with direct types (e.g., 'dose form', 'frequency', 'unit of presentation'). The filtering instruction set enables nuanced corpus pruning by capturing both primary semantic types and their clinically associated aspects.

$$\mathcal{C}_{imc_q}^{\text{filter}} = \{u_{imct} \in \mathcal{C}_{imc_q} \mid \exists \tau \in \mathcal{S}_q : M_{imct}[\tau] > 0\}$$

This guides the metadata filter to select relevant notes, substantially reducing the retrieval corpus while preserving recall of relevant evidence.

3.3.5 Query Entity Extraction and HyDE

To refine filtering, we extract specific entities from query q using MedCAT NER restricted to categories $\mathcal{S}_q$:

$$\mathcal{E}_q^{\text{direct}} = \{e \in \mathcal{E}(q) \mid s(e) \in \mathcal{S}_q\}.$$

In parallel, Hypothetical Document Embeddings (HyDE) (Gao et al., 2023a) generates a hypothetical document d_q from q, that approximates what a relevant clinical note might look like, exposing semantic content and clinical terminology that may not be explicit in the original query phrasing. We use a MedGemma 4B model for this. On this document, we again run NER restricted to $\mathcal{S}_q$, yielding a set of candidate entities

$$\mathcal{E}_q^{\text{hyde, raw}} = \{e \in \mathcal{E}(d_q) \mid s(e) \in \mathcal{S}_q\}.$$

These candidates are then validated against entities observed in the notes corpus, retaining only those that exceed a frequency threshold determined empirically, giving a validated set $\mathcal{E}_q^{\text{hyde}} \subseteq \mathcal{E}_q^{\text{hyde, raw}}$.

The final query entity set is $\mathcal{E}_q = \mathcal{E}_q^{\text{direct}} \cup \mathcal{E}_q^{\text{hyde}}$, which we use to refine filtering:

$$\mathcal{C}_{imc_q}^{\text{filter+}} = \{u_{imct} \in \mathcal{C}_{imc_q}^{\text{filter}} \mid \mathcal{E}_q \cap \mathcal{E}_{imct} \neq \emptyset\}.$$

This entity-level refinement further narrows the filtered corpus, ensuring that subsequent dense retrieval operates over a highly focused and semantically coherent candidate set.

3.3.6 Semantic Retrieval

From the entity-refined filtered set $\mathcal{C}_{imc_q}^{\text{filter+}}$, we perform dense semantic retrieval to obtain an initial candidate set. All chunks across a patient's record are embedded using an encoder for all $u_{imct} \in \bigcup_{i,m,c} \mathcal{C}_{imc}$.

At query time, given query q with embedding $\mathbf{q}$, we compute cosine similarities over the refined filtered corpus and retrieve the top-k' chunks where

$$k' = \min\left(30, \max\left(1, \left\lfloor 0.3 \cdot |\mathcal{C}_{imc_q}^{\text{filter+}}| \right\rfloor\right)\right).$$

This adaptive k' scales with filtered corpus size (taking $\sim$30% of available chunks, capped at 30), ensuring proportional coverage while guaranteeing at least one candidate. These top-k' chunks form the primary evidence set that would typically feed directly into generation in standard RAG frameworks; we instead refine this further with classifier recovery and re-ranking.

3.3.7 Classifier-Based Recall Recovery

Dense retrieval alone may still miss clinically important evidence, especially in long, heterogeneous notes where relevant information can be phrased in diverse ways. To reduce such false negatives, we introduce a recall-oriented classifier that re-examines chunks initially treated as non-relevant.

Let $\mathcal{C}_{imc_q}^{+} \subseteq \mathcal{C}_{imc_q}^{\text{filter+}}$ denote the top-k' chunks from semantic retrieval, with excluded chunks

$$\mathcal{C}_{imc_q}^{\text{neg}} = \mathcal{C}_{imc_q}^{\text{filter+}} \setminus \mathcal{C}_{imc_q}^{+}$$

For the classifier $\phi(q, u_{imct})$, we use a BERT model (Devlin et al., 2019) (`bert-base-uncased`), which demonstrated stronger generalisation than clinical variants in preliminary experiments, to predict chunk-level relevance for $(q, u_{imct}) \in \mathcal{C}_{imc_q}^{\text{neg}}$. We revive chunks exceeding threshold γ (chosen to be 0.8):

$$\widetilde{\mathcal{C}}_{imc_q}^{+} = \mathcal{C}_{imc_q}^{+} \cup \{\, u_{imct} \in \mathcal{C}_{imc_q}^{\text{neg}} \mid \phi(q, u_{imct}) \geq \gamma \,\}$$

Training details, data statistics, and hyper parameters are reported in Appendix A.5.

3.3.8 Re-ranking

After semantic retrieval and classifier recovery, the expanded candidate set $\widetilde{\mathcal{C}}_{imc_q}^{+}$ contains both high-confidence chunks from dense retrieval and revived chunks from classification. To produce a compact, high-quality evidence set for generation, we apply final re-ranking using the `BGE-Reranker` model (Chen et al., 2024) over $\widetilde{\mathcal{C}}_{imc_q}^{+}$:

$$\mathcal{C}_{imc_q}^{\text{final}} =_k \left\{ r(q, u_{imct}) \;\middle|\; u_{imct} \in \widetilde{\mathcal{C}}_{imc_q}^{+} \right\},$$

where $r(q, u_{imct})$ denotes the BGE-Reranker score and k is the final retrieval budget. While semantic retrieval and classifier recovery deliberately cast a wide net to maximise recall, the reranker tightens the candidate set to the final retrieval budget k as described in Section 4.2.2, prioritizing the most query-relevant chunks for the generator.

4 Experiments

4.1 Setup and cohorts

This work evaluates retrieval performance on two complementary cohorts: an LLM-annotated cohort for broad pipeline evaluation, and an expert-annotated cardiovascular cohort constructed alongside clinicians. Both are constructed from de-identified EHR data from MIMIC-III.

4.1.1 LLM-annotated Cohort

This cohort is built by sampling patients at random from MIMIC-III to avoid introducing task-specific or phenotype-specific bias into the evaluation set. For each selected patient, MedGemma 27B generates query–answer-span triplets conditioned on the available record, yielding a patient-specific, visit-specific, and category-specific set of triplets. The spans are then mapped to gold evidence documents in the record, so that each query has both (i) an associated set of relevant documents and (ii) a target answer.

The dataset comprises 1300 query-document-answer triplets derived from 5 patients across 3 visits, with on average 25 documents per query of which 1-2 are gold evidence. For a given patient and visit, queries are generated separately for each note type c rather than over a combined document set. Rather than shallow coverage across many patients, this cohort is designed for depth, extensively interrogating each patient across all visit types and note categories to stress-test the pipeline against the full complexity of real longitudinal records.

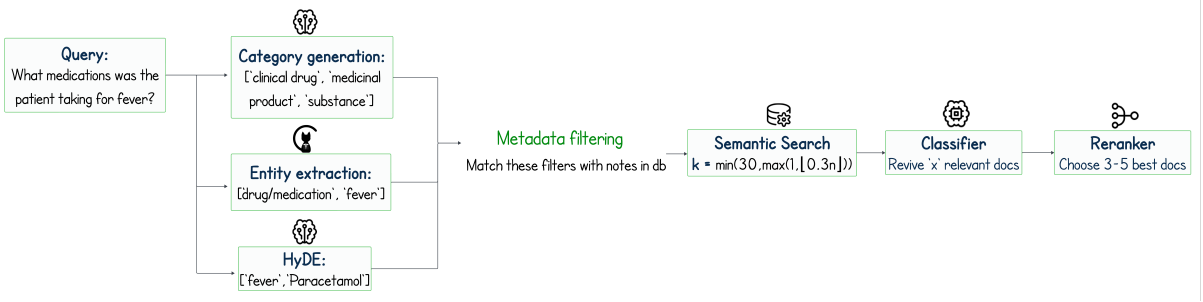


Figure 1: Overview of the four-stage retrieval pipeline

4.1.2 Expert-annotated Cardiovascular Cohort

The second cohort focuses on a clinically coherent subset of cardiovascular patients. Clinical experts first identify relevant patients and then construct 100 query–document mappings, covering typical information needs in cardiovascular medication review. For each patient–visit pair, experts identify relevant documents across all available note types. On average, there are 71 documents for each query, of which about 14 documents are annotated as gold evidence. This cohort therefore serves as the clinical validation benchmark, with high-quality, task-specific annotations against which the pipeline's retrieval performance can be assessed. Unlike existing EHR QA benchmarks that predominantly focus on factoid questions requiring a single supporting document, the average of 14 gold evidence documents per query here reflects the true complexity of real clinical information needs across heterogeneous note types. Appendix A.4 describes the annotation process. A calibration pilot on 30 queries achieved 95.8% inter-annotator agreement, confirming dataset reliability. Across the 100 queries, this dataset amounts to over 7400 document-level relevance judgments, representing a substantial expert annotation effort. This cohort directly reflects the cardiovascular medication review workflow described in Section 6.

Both the cohorts are evaluated under the setting natural to them: the *category-restricted* setting, where retrieval is limited to documents of a specific category c, is applied to the LLM-annotated cohort; and the *visit-level* setting, where retrieval spans all document types within a visit, is applied to the expert-annotated cohort.

4.2 Evaluation Metrics

Notation For a given query, C_{original} denotes all the document chunks in the patient record, C_{filtered} chunks that remain after filtering, G gold chunks associated with the query and C_{revived} the chunks reintroduced by the classifier.

4.2.1 Filtering stage metrics

At the filtering stage, we report:

Corpus ratio The fraction of chunks retained after filtering: $\frac{|C_{\text{filtered}}|}{|C_{\text{original}}|}$

Filtering recall The fraction of gold chunks that remain in the filtered corpus: $\frac{|G \cap C_{\text{filtered}}|}{|G|}$

Revived (classifier configurations only) The fraction of classifier revived chunks that are true gold evidence, reported as 'A/B': $|G \cap C_{\text{revived}}|/|C_{\text{revived}}|$.

4.2.2 Retrieval stage metrics

At the semantic search or reranking stage, operating on the filtered corpus and returning k chunks per query, we report:

Recall@k The fraction of gold chunks appearing among the top-k retrieved chunks.

Averaging All metrics are macro-averaged over queries, so each query contributes equally regardless of the number of associated gold chunks.

Choice of k. We set k=3 for the LLM-annotated cohort (around 25 candidate documents, 1-2 gold) to reflect a low-budget retrieval scenario, and k=20 for the expert-annotated cohort (around 71 candidate documents, 14 gold) to ensure adequate coverage.

4.2.3 Generation stage metrics

We report **ROUGE-L** (Lin, 2004), **BERTScore** (Zhang et al., 2019), and **SARI** (Xu et al., 2016) as generation stage metrics for the LLM-annotated cohort only, as the expert-annotated cohort does not provide reference answers for all queries.

5 Experimental Results

5.1 Baseline Retriever

Our primary baseline is a dense semantic search system that retrieves the top k document chunks per query from the patient corpus, without any additional stages. We use a fixed encoder to obtain embeddings for both queries and document chunks and rank chunks by cosine similarity. This standard RAG retriever configuration serves as the reference system for all subsequent comparisons, representing the most widely deployed retrieval approach in clinical RAG systems.

Baseline configuration The baseline retriever uses the configuration that achieved the best performance across preliminary ablations on chunk strategy, chunk size, and embedding model with full details of the compared methods reported for the LLM-annotated cohort in Appendix A.2.

5.2 LLM-annotated cohort

Table 1 reports retrieval results on the LLM-annotated patient cohort, showing the contribution of each individual component as discussed in Section 5.2.1. The full pipeline shrinks the search space to 78% of original size while boosting recall@3 from 0.60 to 0.73.

5.2.1 Pipeline Ablations

We ablate the pipeline cumulatively, adding metadata filtering, HyDE query expansion, a classifier, and a reranker in sequence, reporting filtering-stage metrics (corpus ratio, filtering recall, revived gold chunks) and recall@3 at each stage.

Metadata filtering Metadata filtering removes more than one quarter of the corpus (0.73), which is attractive from an efficiency standpoint, but the resulting filtering recall of 0.84 is still below the level needed for a reliable end-to-end system. Part of the difficulty is that metadata signals do not fully capture where fine-grained clinical facts are documented, and relevance is highly query dependent, so notes that appear generic or low-priority can still contain critical gold evidence and are occasionally pruned. This motivates the addition of beyond query-aware components.

HyDE HyDE query expansion meaningfully improves both filtering recall and recall@3 when added to the metadata filtered pipeline, raising filtering recall from 0.84 to 0.88 and recall@3 from 0.62 to 0.65. This confirms that queries enriched with clinically plausible terminology surface evidence that the original query phrasing alone misses. However, a filtering recall of 0.88 indicates a non-trivial fraction of gold evidence is missed, motivating a subsequent classifier to recover these.

Classifier The classifier aimed for recall recovery meaningfully boosts filtering recall from 0.88 to 0.94 and recall@3 from 0.65 to 0.69, while reviving 134 gold chunks previously excluded. This demonstrates that learned relevance assessment can correct the over-pruning of borderline cases for generic-looking notes that contain critical clinical facts for which metadata signals alone are insufficient. Together, the three stages build on each other, resulting in substantially stronger evidence recovery.

Reranker The reranker boosts recall@3 from 0.69 to 0.73 by reordering the candidate set, revealing that residual errors reflect fine-grained relevance distinctions among superficially relevant chunks rather than fundamental misses. This confirms the pipeline's strength: each stage corrects a distinct failure mode, culminating in a compact, high-precision evidence set for the generator.

5.3 Expert-annotated Cardiovascular Cohort

The expert-annotated cohort is substantially harder: 71 documents per query versus 25, and 14 gold evidence documents versus 1-2, spanning heterogeneous note types. This is not a controlled retrieval task but a genuine clinical information need, where evidence is distributed and missing any part of it has direct implications for medication review decisions. Despite this complexity, the pipeline delivers a 23% relative recall gain over the dense retrieval baseline, closely matching the +22% observed on the LLM-annotated cohort. The consistency of this gain across two fundamentally different evaluation settings, one broad and automatically annotated and one clinically curated and expert-validated, confirms that the pipeline's improvements are not an artefact of the evaluation conditions but reflect genuine, generalisable evidence recovery. These results demonstrate the pipeline's promise as a practical retrieval tool for real-world clinical use.

5.4 Generation Stage Evaluation

While our primary contribution is retrieval engineering, we include generation metrics to confirm retrieved evidence translates to improved end-to-end QA (Table 2). Using MedGemma 4B

Note: Each approach builds cumulatively on all prior stages

Table 1: Ablated retrieval results under natural configurations

Approach	Corpus Ratio	Filtering Recall	Revived	Recall@k
LLM-annotated cohort (category-restricted)				
Baseline	1.00	–	–	0.60
Metadata filtering	**0.73**	0.84	–	0.62
HyDE	0.75	0.88	–	0.65
Classifier	0.78	**0.94**	134/514	0.69
Reranker (Full Pipeline)	0.78	**0.94**	**376/1369**	**0.73**
Expert-annotated cohort (visit-level)				
Baseline	1.00	–	–	0.47
Metadata filtering	**0.65**	0.77	–	0.52
HyDE	0.66	0.80	–	0.52
Classifier	0.68	**0.83**	21/71	0.56
Reranker (Full Pipeline)	0.68	**0.83**	**38/91**	**0.58**

Table 2: Generation stage metrics

Metric	Baseline	Full pipeline
ROUGE-L	0.362	0.412
BERTScore	0.469	0.533
SARI	74.92	82.45

conditioned on top-k retrieved chunks, the full pipeline improves all generation metrics over baseline retrieval: ROUGE-L: +14% relative gain, BERTScore: +14% relative gain and SARI: +10% relative gain. These consistent gains across lexical overlap (ROUGE-L), semantic similarity (BERTScore), and system output quality (SARI) validate that improvements propagate to answer quality. Our focus on deterministic retrieval provides the strong evidence foundation on which generation quality directly depends.

6 Clinical Workflows

The pipeline is designed for real-world clinical deployment, with two immediate target workflows:

- *For clinical risk scoring*, such as CHA_2DS_2-VASc score (Lip et al., 2010) computation, where each variable must be evidenced by specific documentation across heterogeneous note types, high recall is the critical requirement, as missing a single signal can produce an incorrect risk score. The pipeline's multi-stage design ensures that evidence scattered across note types is systematically recovered, reducing the risk of incorrect risk scoring due to missed clinical signals.
- *For cardiovascular medication review*, clinicians must synthesize evidence distributed across discharge summaries, nursing notes, ECG reports, and echo reports. The pipeline surfaces this evidence in response to natural language queries, enabling clinicians to interrogate a patient's longitudinal record directly. The pipeline's deterministic, stage-wise design ensures every retrieval decision remains traceable and auditable, properties as important as performance in settings where the provenance of evidence directly affects clinical accountability.

7 Conclusion

Reliable clinical question answering over EHRs depends fundamentally on retrieval quality, yet the retrieval component has remained the least scrutinized part of clinical RAG systems. This work addresses that gap directly. By decomposing retrieval into four distinct, instrumented stages, the pipeline makes evidence recovery measurable and auditable at every step, achieving 22-23% relative recall gain over a strong dense retrieval baseline across cohorts. The consistency of this gain across two evaluation settings of fundamentally different complexity suggests genuine, generalisable improvements in evidence recovery rather than artefacts of evaluation design. Our work delivers a retrieval system that clinicians and researchers can rely on, inspect, understand, and build upon for real-world clinical use.

Limitations and Future Work

The LLM-annotated cohort was designed to prioritise depth over breadth, extensively interrogating five patients across all visit types, note categories, and query distributions to stress-test the pipeline against the full complexity of real longitudinal records. While this design choice ensures thorough within-patient evaluation, future work should extend coverage to a broader patient population to assess robustness across variation in documentation style and clinical complexity.

Similarly, evaluation is currently restricted to MIMIC-III, a single US medical centre; applying the pipeline to other EHR systems, clinical domains, languages, and documentation cultures is a natural and important next step.

Generation metrics are reported for MedGemma 4B only, and evaluating downstream answer quality under a broader range of generators is a straightforward extension. The pipeline's recall ceiling is tied to MedCAT's NER performance, and future work could explore more clinical NER systems or LLM-based entity extraction.

The classifier is trained on a heavily imbalanced dataset reflecting realistic clinical sparsity, and evaluating threshold robustness across note types and query distributions beyond those seen in training remains an open question. Finally, the multi-stage design introduces additional latency over a standard dense retrieval baseline; detailed runtime metrics and latency characterization for time-sensitive clinical deployment will be reported in future work.

Acknowledgments

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SA, TS, RD are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RD is also supported by The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics

This work uses de-identified patient data from MIMIC-III under standard data use agreements. All clinical annotations were performed by qualified experts following structured guidelines. The pipeline is intended to support clinical decision-making, not replace clinical expertise, and retrieval outputs should always be reviewed by a clinician before informing patient care. The deterministic, auditable design of the pipeline is a deliberate ethical choice, ensuring that retrieval decisions remain transparent and traceable rather than opaque. We acknowledge the risk of retrieval bias, and note that evaluation on a broader and more diverse patient cohort remains important future work.

References

Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. 2024. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of biomedical informatics*, 156:104662.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alan Aronson. 2010. An overview of metamap: Historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Yu Gu, Robert Tinn, Hao Cheng, and et al. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Jordan R Hill, Shyam Visweswaran, Xia Ning, and Titus K Schleyer. 2021. Use, impact, weaknesses, and advanced features of search functions for clinical use in electronic health records: a scoping review. *Applied Clinical Informatics*, 12(03):417–428.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, and et al. 2023. Active retrieval augmented generation. In *EMNLP*.

Di Jin, Eileen Pan, Nathalie Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, and et al. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, David Vickers, and Liaqat Butt. 2015. Information retrieval as a tool for clinical decision support: A systematic review. *JAMIA*, 22(4):799–811.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, and et al. 2021a. Medcat: The medical concept annotation toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021b. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, and et al. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2023. Do we still need clinical language models? *EMNLP*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Ying Li, Hojjat Salmasian, Rave Harpaz, Herbert Chase, and Carol Friedman. 2011. Determining the reasons for medication prescriptions in the ehr using knowledge and natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2011, page 768.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

NHS. 2023. Purpose of the gp electronic health record. Accessed on March 18, 2024.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and 1 others. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Jyotishman Pathak, Abel N Kho, and Joshua C Denny. 2013. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2):e206–e211.

Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Summits on Translational Science Proceedings*, 2014:218.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. In *The... text REtrieval conference: TREC. Text REtrieval Conference*, volume 26, pages https–trec.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Guergana Savova, James Masanz, Philip Ogren, and et al. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes). *JAMIA*, 17(5):507–513.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, and et al. 2023. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*.

Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

João Figueira Silva and Sérgio Matos. 2022. Modelling patient trajectories using multimodal information. *Journal of biomedical informatics*, 134:104195.

Karan Singhal, Shekoofeh Azizi, Tao Tu, and et al. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: a literature review. *Journal of healthcare informatics research*, 8(2):313–352.

SNOMED. Snomed international. Accessed on March 27, 2024.

Irena Spasic, Goran Nenadic, and 1 others. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval

with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Shunyu Yao, Jeffrey Zhao, Dian Yu, and et al. 2023. React: Synergizing reasoning and acting in language models. In *ICLR*.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Rui Zhang, Serguei VS Pakhomov, Elliot G Arsoniadis, Janet T Lee, Yan Wang, and Genevieve B Melton. 2017. Detecting clinically relevant new information in clinical notes across specialties and settings. *BMC medical informatics and decision making*, 17(Suppl 2):68.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Algorithm for Chunking

Algorithm 1 Chunking clinical notes

Require: Notes grouped by patient, admission, and category
Require: Minimum chunk size L_{min}, maximum chunk size L_{max}
Ensure: A list of text chunks with timestamps

```
Initialize empty list of chunks C
for each patient–admission–category group G do
    Sort notes in G chronologically
    Start an empty chunk buffer B, length L ← 0
    for each note in order do
        Let ℓ be the length of this note
        if ℓ is very large then
            Split the note into smaller pieces using recursive splitter
            for each piece do
                if adding it to B would exceed L_max then
                    Save current chunk to C and start a new one
                end if
                Add piece to current chunk and record its timestamp
            end for
        else if adding the note would exceed L_max then
            Save current chunk to C
            Start a new chunk with this note
        else
            Add the note to the current chunk
        end if
    end for
    if the final chunk buffer is not empty then
        Save it to C
    end if
end for
return all chunks C
```

A.2 Baseline Design Choices

Before introducing the full multi-stage pipeline, we first investigate two design choices for the baseline retriever: chunk size and embedding model. These experiments allow us to select a strong and well-justified baseline configuration, which we then use consistently in the main results and pipeline ablations.

A.2.1 Chunking Strategy and Size

We conduct an ablation study on different chunking strategies and chunk sizes to determine the most effective configuration for the baseline retriever. The chunking strategies compared are:

- **Recursive chunking:** Fixed-length recursive segmentation with partial overlap between segments.
- **Smart structural chunking (proposed):** A structure-aware approach that respects document boundaries and clinical section headers while maintaining approximately uniform chunk length.

For each strategy, we experimented with multiple target chunk sizes (approximately 400, 1200, and 1800 characters) and evaluated recall@3 using the dense retriever. The results, summarised in Table 3, show that the smart structural chunking strategy consistently outperforms purely length-based segmentation. A chunk size of around *1000* offered the best trade-off between retrieval recall and corpus compactness and was selected for all subsequent experiments.

We note that more sophisticated approaches such as semantic chunking or sentence-level segmentation were not explored in this work. In the clinical domain, such methods often disrupt the logical flow of medical narratives, where the order of statements, temporal markers, and contextual dependency carry significant meaning. To preserve the semantic continuity and clinical validity of the text, we retain structure-aware, non-semantic chunking methods.

A.3 Embedding Model

Using the optimal chunking configuration identified above, we compared several embedding models for the baseline retriever, including `nomic-ai/nomic-embed-text-v1.5` (Nussbaum et al., 2024), `intfloat/e5-base-v2` (Wang et al., 2022), `mixedbread-ai/mxbai-embed-large-v1` (Lee et al., 2024), `emilyalsentzer/Bio_ClinicalBERT` (Alsentzer et al., 2019) and `sentence-transformers/all-MiniLM-L6-v2` (Thakur et al., 2021). Each model was used to encode both query and chunk embeddings,

and performance was measured using recall@3. The results, shown in Table 4, demonstrate that `llm-embedder` provides the highest recall while maintaining computational efficiency. It was therefore selected as the default embedding model for both the baseline and all full-pipeline experiments.

Strategy	Chunk size	Recall@3
Recursive chunking	400	0.41
	1200	0.53
	1800	0.56
Smart chunking	(200, 500)	**0.45**
	(900, 1500)	**0.60**
	(1700, 2200)	**0.59**

Table 3: Chunk size ablation

Model	Recall@3
llm-embedder	**0.6**
nomic-embed-text-v1.5	0.59
e5-base-v2	0.58
mxbai-embed-large-v1	0.59
Bio_ClinicalBERT	0.42
all-MiniLM-L6-v2	0.55

Table 4: Embedding model ablation

A.4 Expert Annotation Process

Two clinical experts performed annotation for the cardiovascular cohort using a structured process:

Annotation guidelines We created a comprehensive annotation document detailing:

- **Task definition**: Identify all documents containing evidence for clinical information needs
- **Relevance criteria**: Direct mentions, inferable evidence, multi-hop reasoning requirements
- **Annotation procedure**: Consensus resolution, temporal judgment guidelines

Annotation tool We built a custom web interface to streamline the process, featuring:

- Side-by-side patient record browser (visit-organized notes by category)
- Query-document relevance marking
- Export to structured gold format for evaluation

Pilot and calibration An initial pilot on the same 30 queries achieved 95.8% inter-annotator agreement. A calibration meeting refined edge cases:

- Reasoning "hops": single-hop (direct) vs. multi-hop (synthesis across mentions)
- Temporal ambiguity in nursing notes (past vs. present conditions)
- Cross-document evidence distribution across note types

Remaining 70 queries were split among the annotators.

A.5 Classifier Training Details

Training data consists of query-chunk pairs generated using Gemma 12B (Team et al., 2025) over a held-out set of MIMIC-III patients with no overlap with either evaluation cohort. Chunks containing gold evidence serve as positive examples and non-relevant chunks within the same patient-visit-category serve as negatives, yielding 16894 training pairs in total with 2238 positive and 14656 negative, reflecting the realistic sparsity of relevant evidence in clinical notes. The classifier is trained with binary cross-entropy loss with class weights for 10 epochs, with learning rate 5×10^{-5} and weight decay 0.01.

Interpretable ICD Code Classification with Faithful Sentence Extraction

Yichen Wang[†*], **Lian Hong**[†*], **Masato Mizogaki**[‡‖*], **Shunnosuke Umeda**[†], **Toshimune Kenmotsu**[†], **Akihiro Tamura**[§] and **Daniel Andrade**[†¶]
[†] Hiroshima University, Japan, andrade@hiroshima-u.ac.jp
[‡] Bourbon Corporation, Japan
[§] Doshisha University, Japan

Abstract

Transformer-based models such as PLM-CA achieve strong performance for automatic ICD coding, but their attention weights do not provide faithful explanations of their predictions. This is a major limitation for electronic medical records, where users often need concise and trustworthy evidence for each assigned code. To address this issue, we jointly train a sentence extractor and an ICD code classifier such that predictions are based only on the extracted sentences. As a result, the extracted sentences serve as faithful rationales for each predicted code and substantially reduce the effort required to inspect long medical records. Experiments on MIMIC-III show that our method approaches the performance of a transformer baseline that processes the full record while using only a small fraction of the document.

1 Introduction

The International Classification of Diseases (ICD), defined by the World Health Organization, helps match patients to appropriate treatments.[1] However, the manual assignment of ICD codes is labor-intensive and requires substantial expertise. Consequently, automatic ICD coding from electronic medical records has been the focus of extensive research (Dong et al., 2022; Edin et al., 2023). Recent advances in pretrained transformer-based models have further improved ICD coding performance (Edin et al., 2023; Aden et al., 2024). In particular, Edin et al. (2024) proposed PLM-CA, a model pretrained on medical data that achieves state-of-the-art classification performance.

Unfortunately, attention mechanisms in transformer-based classification models do not provide faithful explanations for ICD code predictions (Bastings and Filippova, 2020). To address this limitation, we jointly train a sentence extractor and an ICD code classifier such that the classifier relies only on the selected sentences. Following Lyu et al. (2024), we refer to such a classifier as faithful.

Our working hypothesis is that a few key sentences, also referred to as rationales (Ehsan et al., 2019), can be sufficient to correctly judge each ICD code of a medical record. As a consequence, we propose the following two step method: (1) Select a subset of relevant sentences $\hat{S}$ from the document. (2) Predict the ICD code with an interpretable classifier using only $\hat{S}$.

With the requirement of an independent sentence selection in step (1), we prove the faithfulness of the classifier, and show that the parameters of the models in step (1) and (2) can be jointly trained (see Section 3). Furthermore, our experiments using the MIMIC-III dataset (Johnson et al., 2016) show that the proposed method can successfully identify a small number of relevant sentences that are critical for the correct ICD code classification (see Section 4).

2 Background and Related Works

ICD code prediction for electronic medical records is often formulated as a (multi-label) classification task (Edin et al., 2023). ICD code prediction is challenging due to the fact that there are several thousand codes, where many are rare and *not* mutually exclusive. As a result, many works have focused on increasing prediction accuracy (Edin et al., 2023; Dong et al., 2022).

[*] Equal contribution.
[‖] Work done at Hiroshima University during bachelor, not associated with Bourbon Corporation.
[¶] Corresponding author.
[1] https://icd.who.int/

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 679–686
July 3-4, 2026 ©2026 Association for Computational Linguistics

Transformer-based models are often among the best performing models (Edin et al., 2023; Ravichandran et al., 2024). In particular, PLM-CA is considered to be one of the best models for ICD code prediction. The PLM-CA model, proposed by Edin et al. (2024), is an improved version of the PLM-ICD model of Huang et al. (2022), with a multi-head attention mechanism for ICD codes and RoBERTa pretrained on medical data (Lewis et al., 2020). Though, it is tempting to use a model's attention-mechanism for producing explanations of the prediction (Ravichandran et al., 2024; Mullenbach et al., 2018), these explanations are not faithful (Bastings and Filippova, 2020; Lyu et al., 2024). In particular, there are no theoretical guarantees that removing any non-highlighted text will not change the outcome of the classifier. For example, Feng et al. (2018) demonstrate that using attention scores for rationale extraction can lead to unintuitive behavior that is difficult to interpret.

Using sentences as rationales has also been considered (Herrewijnen et al., 2021; MISAWA et al., 2023), but those methods use information from the entire document for sentence selection, which violates faithfulness, as we describe in Proposition 1. Other methods require manually annotated rationales that are expensive for specialized medical domains (Chan et al., 2022).

Zhou et al. (2021) and Douglas et al. (2025) show that only a small fraction of a discharge summary is relevant for ICD coding. In particular, Douglas et al. (2025) propose a filtering step before applying PLM-CA to retain only clinically relevant text spans. However, this filtering step requires a named entity recognizer and an assertion classifier, both of which rely on fine-grained annotations. To improve interpretability, they also use AttnInGrad (Edin et al., 2024), which still does not guarantee faithfulness.

3 Proposed Method

Our goal is a classification method that bases its prediction result only on a small subset of rationales from a long document. Here, we consider each sentence as a potential rationale.

Since many ICD codes are not mutually exclusive, we consider a multi-label classification model. Let C denote the number of classes (=ICD codes). Assuming a document contains the sentences $\{s_1, s_2, \ldots, s_m\}$, we denote set of their indices by $S := \{1, 2, \ldots, m\}$. Given a constraint $0 < k < m$ on the number of rationales, for each class $c \in \{1, 2, \ldots, C\}$, we proceed as follows:

1. Using some function $h_{\boldsymbol{\eta}}$, we embed each sentence s_i *independently* into a vector representation $\mathbf{x}_i \in \mathbb{R}^H$.

2. We score each sentence for class c using some function $f^c_{\boldsymbol{\theta}} : \mathbb{R}^H \to \mathbb{R}$, and select the sentences with the top-k scores, i.e. $\hat{S}^c := \{i \in S \mid f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(k)})\}$, where

$$f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(1)}) \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(2)}) \ldots \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(k)}) \ldots \geq f^c_{\boldsymbol{\theta}}(\mathbf{x}_{(m)}) .$$

3. We pool the sentence embeddings with the top-k scores and pass the resulting vector again to $f^c_{\boldsymbol{\theta}}$, which return value is interpreted as the logit for class c. That means the probability that the document is assigned ICD code c is given by

$$p(\mathbf{y}_c = 1|S) = \text{sigmoid}(f^c_{\boldsymbol{\theta}}(pool(\hat{S}^c)) , \quad (1)$$

where $\mathbf{y} \in \{0, 1\}^C$ is the class label vector of the corresponding document.

The parameters of the sentence embedder and classifier, namely $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, are jointly trained.

By construction, the proposed method has the following important property.

Proposition 1 *Removing from the document any sentence from $S \backslash \hat{S}^c$ leaves the classification result for class c unchanged.*

Proposition 1 is a basic requirement for a classifier with faithful explanations. However, transformer-like attention mechanisms do not fulfill this requirement, since the attention score for a sentence (or token) itself depends on all sentences.

Finally, we note that the proposed method (and Proposition 1) have the subtle requirement that $m > k$. Since our focus is on classification of long medical records this is not a limitation. However, we note that as an alternative to taking the top k sentences, we could introduce a threshold τ, and then use all sentences

with $f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \tau$. This still fulfills Proposition 1, though, this has the disadvantage that for some documents this may lead to a large set $\hat{S}^c$, where many rationales have overlapping information content.

3.1 Joint Sentence Extraction and Classification with LLM

Below, we describe a concrete instance of the proposed method. First, using a large language model (LLM) denoted by $h_{\boldsymbol{\eta}}$, we encode each sentence s_i as a representation $\mathbf{x}_i \in \mathbb{R}^H$, i.e., $\mathbf{x}_i := h_{\boldsymbol{\eta}}(s_i)$. Because of the large number of ICD codes, we use a single LLM $h_{\boldsymbol{\eta}}$ for all classes. Next, let $W \in \mathbb{R}^{C \times H}$ and $\mathbf{b} \in \mathbb{R}^C$ denote the parameters of a linear multi-label classifier over all ICD codes. For each sentence-label pair, we compute

$$f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) = \mathbf{w}_c^\top \mathbf{x}_i + b_c \,,$$

where $\mathbf{w}_c^\top$ is the c-th row of W.

For the pooling function in Equation (1), we use the mean, i.e.,

$$pool(\hat{S}^c) := \frac{1}{|\hat{S}^c|} \sum_{i \in \hat{S}^c} \mathbf{x}_i \,,$$

where $|\hat{S}^c| = k$.

The parameters of the LLM and the classifier, namely $\boldsymbol{\eta}$, W, and $\mathbf{b}$, are trained jointly. We train the model in two stages. In Stage 1, we use all sentences in S to obtain a stable initialization. In Stage 2, we switch to the top-k selection mechanism described above. Since ICD coding is a multi-label task, we optimize binary cross-entropy loss over sigmoid outputs. We denote the resulting method by LLM-SeparateTopK.

3.1.1 Common Sentence Extraction

Note that the method described above extracts k rationales for each ICD code. Alternatively, we also explore a variation where we extract in *total* only k sentences that are used for classification. For that we modify step 2 (Section 3), by determining the set of selected sentences $\hat{S}$ as follows. Define $\hat{S} := \{i \in S \,|\, \alpha_i \geq \alpha_{(k)}\}$, where

$$\alpha_{(1)} \geq \alpha_{(2)} \ldots \geq \alpha_{(k)} \ldots \geq \alpha_{(m)} \,,$$

and α_i is the aggregated positive contribution of sentence i across all labels:

$$\alpha_i := \sum_{c=1}^{C} \mathrm{ReLU}(f^c_{\boldsymbol{\theta}}(\mathbf{x}_i) - b_c) \,,$$

where we subtract the bias term to ensure that each class is given the same weight. We denote the resulting method as LLM-GlobalTopK.

3.2 Bag-of-words Model

Alternatively to the usage of an LLM, we also consider a Bag-of-words (BoW) model for $\mathbf{x}_i$, where we use all unigrams, bi-grams, and tri-grams that are contained in at least 3 sentences. Note that in this case the embedding function h has no trainable parameters. We denote the corresponding methods as BoW-SeparateTopK and BoW-GlobalTopK.

4 Experiments

4.1 Dataset and Preprocessing

For our experiments, we use the discharge summaries of the Beth Israel Deaconess Medical Center collected during 2001 to 2012, which are part of the MIMIC-III dataset (Johnson et al., 2016). For the preprocessing and train/validation/test splits we use MIMIC-III clean as in (Edin et al., 2023).[2] Each document is split into sentences using the NLTK library (Bird et al., 2009) subject to some post-processing described in Appendix A. Note that the resulting median number of sentences per document is 141.

4.2 Baselines

A simple interpretable baseline is a bag-of-words model that uses the whole document (BoW-All). Furthermore, as a rationale extraction baseline, we compare against a method that uses the PLM-CA model to identify important sentences with Integrated Gradients (Sundararajan et al., 2017), and then retrains PLM-CA. We call the resulting method IG-GlobalTopK and provide details in Appendix C. We also compare against methods that simply extract the first k sentences of the document

[2] See https://github.com/JoakimEdin/medical-coding-reproducibility and https://github.com/JoakimEdin/explainable-medical-coding/tree/main/data/splits.

	Classification				Ranking		
	AUC-ROC		F1		Precision@r		MAP
	Micro	Macro	Micro	Macro	8	15	
Proposed methods							
LLM-SeparateTopK (k = 10)	**98.2**	**94.2**	**51.7**	**18.3**	**65.0**	**50.2**	**55.5**
LLM-GlobalTopK (k = 10)	97.1	89.8	46.0	13.5	59.4	44.6	48.4
BoW-GlobalTopK (k = 10)	94.6	78.1	25.7	2.9	49.9	37.8	39.0
BoW-SeparateTopK (k = 10)	94.0	75.3	38.5	5.8	58.8	44.5	47.2
Baseline methods							
IG-GlobalTopK (k = 10)	96.1	86.1	25.6	1.9	45.5	34.7	35.4
BoW-Head (k = 10)	93.4	74.2	21.1	2.09	45.4	34.4	35.0
BoW-Tail (k = 10)	91.3	66.6	13.4	0.92	36.9	28.5	27.6
BoW-All	94.3	76.1	35.5	5.1	57.6	43.5	46.6
LLM-Head (k = 10)	96.0	86.7	35.6	7.4	45.9	35.0	36.0
LLM-Tail (k = 10)	94.2	80.7	29.1	3.8	36.8	28.6	28.0
LLM-All	95.4	85.0	38.6	10.2	52.2	39.2	41.5
PLM-CA	**98.9**	**95.7**	**59.8**	**29.0**	**72.3**	**56.7**	**65.0**

Table 1: Comparison of baseline and proposed methods for ICD code prediction in terms of micro/macro AUC-ROC, micro/macro F1, MAP, and Precision@r (r = 8, 15). MAP: mean average precision. Best results highlighted in bold, for proposed and baseline methods, respectively.

(i.e., the head sentences) or the last k sentences (i.e., the tail sentences), which we denote as BoW-Head and BoW-Tail, respectively. Analogously, we compare to a method that pools all sentence embeddings (LLM-All), the first k sentence embeddings (LLM-Head), or the last k ones (LLM-Tail). Finally, for reference, we also report the results of the original PLM-CA model, which uses the whole document (Douglas et al., 2025).

4.3 Results

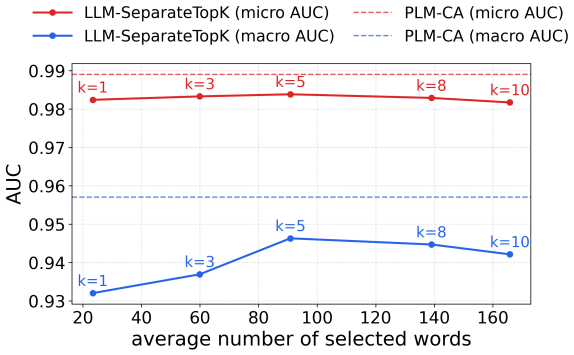


Figure 1: Proposed method's AUC scores for different values of k.

For the proposed methods and corresponding baselines, we set $k = 10$, which reduces most documents by more than 90%. We evaluate all methods on all 3,681 ICD codes occurring in the training and test splits, and report micro/macro AUC-ROC, micro/macro F1, mean average precision (MAP), and Precision@r in

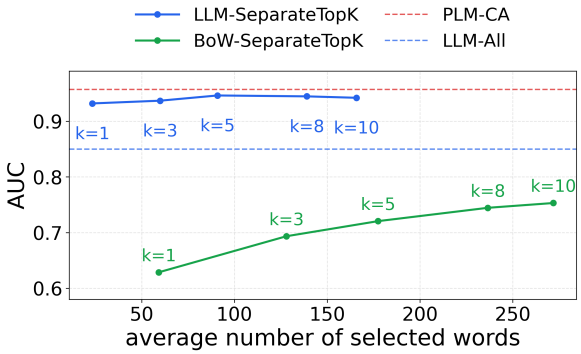


Figure 2: Comparison of different method's macro AUC scores for different values of k.

Table 1. Figure 1 shows the AUC scores for different values of k for the proposed method.

Our results show that the proposed classifier remains accurate while using only a small subset of each document. In particular, for $k = 10$, LLM-SeparateTopK performs better than all baselines in Table 1, including the rationale-based baseline IG-GlobalTopK as well as naive head/tail extraction methods. Although full-document PLM-CA remains a strong reference point, the proposed method often approaches its performance while providing faithful rationale extraction. Table 2 shows one example of the extracted sentences.

Finally, in Figure 2, we also show the macro AUC of the proposed method LLM-SeparateTopK for different number of selected sentences k. Interestingly, we see that the pro-

posed method performs *better* than LLM-All, suggesting that selecting only few sentences has the additional benefit of removing noise.

Rank	Extracted Sentence
1	"discharge diagnosis angioedema"
2	"brief hospital course angioedema followed by dr"
3	"you were admitted to the hospital with throat swelling which was an exacerbation of your chronic angioedema"

Table 2: Top three extracted sentences produced by LLM-SeparateTopK ($k = 10$) for ICD code 995.1 (Angioneurotic edema) from a discharge summary.

5 Conclusions

In this paper, we proposed a provably faithful approach to rationale extraction for ICD code prediction. Experiments on MIMIC-III showed that our method can extract a small number of sentences *without any sentence- or token-level annotations* while achieving competitive coding performance, even compared with methods that process the full discharge summary. Overall, our results suggest that faithful sentence-level rationales are a promising step toward more transparent and clinically useful medical coding systems.

As future work, we plan to evaluate the extracted rationales against human-annotated evidence resources such as MDACE (Cheng et al., 2023), which would also allow us to assess plausibility.

Limitations

The current method judges the importance of each sentence independently of context. While this is an important requirement for the faithfulness, as described in Proposition 1, this might lead to wrongly excluding short sentences that are important when judged in the context of its surrounding sentences.

References

Ilyas Aden, Christopher HT Child, and Constantino Carlos Reyes-Aldasoro. 2024. International classification of diseases prediction from mimiic-iii clinical text using pre-trained clinicalbert and nlp deep learning models achieving state of the art. *Big Data and Cognitive Computing*, 8(5):47.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*, pages 2867–2889. PMLR.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. 2022. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159.

James C. Douglas, Yidong Gan, Ben Hachey, and Jonathan K. Kummerfeld. 2025. Less is more: Explainable and efficient ICD code prediction with clinical entities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30835–30847, Vienna, Austria. Association for Computational Linguistics.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on mimic-iii and mimic-iv: a critical review and replicability study. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2572–2582.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th international*

conference on intelligent user interfaces, pages 263–274.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Elize Herrewijnen, Dong Nguyen, Jelte Mense, Floris Bex, and 1 others. 2021. Machine-annotated rationales: faithfully explaining text classification. In *Proceedings for the Explainable Agency in AI Workshop at the 35th AAAI Conference on Artificial Intelligence (Washington DC: AAAI Press)*, pages 11–18.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Shotaro MISAWA, Taiki FURUKAWA, Shintaro OYAMA, Ryuji KANO, Hirokazu YARIMIZU, Tomoki TANIGUCHI, Kohei ONODA, Kikue SATO, and Yoshimune SHIRATORI. 2023. Sentence extraction using outcome prediction model trained from clinical data. *Proceedings of the Annual Conference of JSAI*, JSAI2023:3Xin404–3Xin404.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Roland Roller, and Sebastian Möller. 2024. XAI for better exploitation of text in medical decision support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513, Bangkok, Thailand. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, Online.

A Sentence Splitting

First, each document is split into sentences using the sent_tokenize function[3] of the NLTK library (Bird et al., 2009). Afterwards, the following post-processing steps are used to reduce conflation and improve comprehensiveness:

1. If there are two or more line breaks, split the sentence before and after the line break.

2. If there are two or more spaces, split the sentence before and after the spaces.

3. Replace the line break characters with spaces.

4. Remove spaces at the beginning and end of the sentence.

5. If each sentence ends with a ":", join that sentence to the next sentence with a space.

The resulting median number of sentences (and IQR) per document is 141 (102-189). The median number of words (and IQR) in a sentence are 8 (3-13). The median number of words (and IQR) in a document is 1,375 (965-1,900).

[3] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

B Training Details

For LLM-SeparateTopK, we trained the model for up to 9 epochs and used early stopping with a patience of 3 epochs according to the validation loss. The learning rates for the pretrained encoder and the linear classifier were set to 3×10^{-6} and 5×10^{-5}, respectively. Owing to the memory requirement of the model, we used a mini-batch size of 1 document and accumulated gradients for 2 steps, yielding an effective batch size of 2. No explicit regularization was used; instead, early stopping was employed to mitigate overfitting. All experiments were run on an NVIDIA Blackwell GPU, and each epoch took around 2 hours.

C Sentence-level Top-k Selection via Integrated Gradients (IG-GlobalTopK)

As a rationale-extraction baseline, we use Integrated Gradients (IG) (Sundararajan et al., 2017) to score sentences and then retrain PLM-CA on the extracted sentences only.

Given a document with S sentences, where each sentence s_i consists of a set of tokens T_i, we first compute the attribution of each token. Let $F : \mathbb{R}^n \to \mathbb{R}$ be the model's loss function (Binary Cross-Entropy), $\mathbf{u} \in \mathbb{R}^n$ be the vector of all tokens in the document, and $\mathbf{u}' \in \mathbb{R}^n$ be a baseline consisting of PAD tokens. The IG score for the j-th token is defined as:

$$\mathrm{IG}_j(\mathbf{u}) = (u_j - u'_j) \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{u}' + \alpha(\mathbf{u} - \mathbf{u}'))}{\partial u_j} d\alpha.$$

To obtain a document-level importance score at the sentence level, we aggregate the token-wise attributions within each sentence by calculating their mean:

$$\mathrm{Score}(s_i) = \frac{1}{|T_i|} \sum_{j \in T_i} \mathrm{IG}_j.$$

Similar to the top-k selection of our proposed method, we then define

$$\hat{S}_{\mathrm{IG}} := \{ i \in S \,|\, \mathrm{Score}(s_i) \geq \mathrm{Score}(s_{(k)}) \},$$

where

$$\mathrm{Score}(s_{(1)}) \geq \mathrm{Score}(s_{(2)}) \geq \ldots \geq \mathrm{Score}(s_{(m)}).$$

We then retrain PLM-CA using only $\hat{S}_{\mathrm{IG}}$, while preserving the original sentence order.

To prevent discrepancies between training and testing, we use the following retraining procedure: we partition the training set into five folds: 4/5 are used for training and 1/5 for IG-based sentence selection, and repeat this 5 times. Finally, we train a new classifier $\mathcal{C}$ on the extracted sentences from the whole training data.

The resulting method, denoted by IG-GlobalTopK, uses at test time classifier $\mathcal{C}$ applied to extracted sentences only. However, note that IG-GlobalTopK is not strictly faithful: a sentence s_i that was *not selected* could have influenced the choice of $\hat{S}_{\mathrm{IG}}$. This in consequence, allows sentence s_i to influence indirectly the final classification result.

D Example of Extracted Sentences

We provide additional examples here, including both successful and failure cases for LLM-SeparateTopK (Tables 3 and 4), as well as a successful case for LLM-GlobalTopK (Table 5).

ICD Code	493.90 long description: Asthma, unspecified type, unspecified
Rank	**Extracted Sentence**
1	"asthma continued flovent prn"
2	"asthma"
3	"fluticasone mcg actuation aerosol sig one puff inhalation hospital1 times a day as needed for shortness of breath or wheezing"

Table 3: An additional successful example of extracted sentences produced by the proposed LLM-SeparateTopK ($k = 10$) for ICD code 493.90 from one discharge summary (admission id = 179767). The selected sentences directly indicate the target respiratory condition and its related medication use.

ICD Code	886.0 long description: Traumatic amputation of other finger(s), complete, without mention of complication
Rank	**Extracted Sentence**
1	"the patient underwent completion amputation of the left ring finger repair of digital ulnar and radial nerve to the middle finger index finger and thumb repair of the fpl ftp of the middle finger and fts ftp of the index finger repair of the epl to the thumb and k wire fixation of the metacarpal of index and phalanx of the first finger"
2	"history of present illness the patient arrived from another hospital in hospital3 to the emergency room with a complete amputation of his thumb index long and ring fingers of his left hand"
3	"discharge diagnosis as described above amputation of thumb index long and ring fingers reattachment of thumb index and long fingers and revision amputation of ring finger"

Table 4: A representative failure example of extracted sentences produced by the proposed LLM-SeparateTopK ($k = 10$) for ICD code 886.0 from one discharge summary (admission id = 174415). Although the selected sentences are clinically plausible and strongly related to traumatic finger injury, the model incorrectly assigns this code.

ICD Code	995.1 and 493.90
Rank	**Extracted Sentence**
1	"asthma"
2	"chronic urticaria and angioedema since c section done on for failure to progress after hour labor"
3	"brief hospital course angioedema followed by dr"
4	"discharge diagnosis angioedema"
5	"asthma continued flovent prn"
6	"physical exam t bp hr rr o2 sat ra wt kg gen nad speaking in full sentences without sob no audible stridor heent moon facies mmm angioedema no edema noted in posterior oropharynx and posterior oropharynx clearly visualized behind tongue neck supple cv rrr s1 split s2 ii vi systolic murmur over lsb lungs cta b l without wheezing rhonchi or rales abd soft obese nt nd normoactive bs ext no le edema wwp neuro aao x skin no rashes noted"
7	"history of present illness yo f with h o idiopathic chronic urticaria and angioedema requiring intubation during previous admission asthma and seasonal allergies who presents with tongue swelling x hrs"
8	"medications on admission epipen prn loratadine mg daily ranitidine mg qam doxepin mg qhs benadryl mg q6h prn calcium mg daily vit d units daily ativan mg qhs flovent mcg puff hospital1 prn for angioedema"
9	"first name4 namepattern1 last name namepattern1 as outpt for idiopathic chronic urticaria and angioedema with multiple admissions in past year for angioedema"
10	"otherwise no history of related urticaria angioedema"

Table 5: An additional successful example of extracted sentences produced by the proposed LLM-GlobalTopK ($k = 10$) for all ICD code from one discharge summary (admission id = 179767). The selected sentences directly indicate the target respiratory condition and its related medication use.

Evaluating LLM-as-a-Judge for Medical Term Simplification

Ioana Buhnila[1,2] **Aman Sinha**[1,3,4] **Rohit Agarwal**[5]
Dilip K. Prasad[5] **Mathieu Constant**[1]

[1]ATILF, University of Lorraine - CNRS, France
[2]Center for Data Science in Humanities, Chosun University, South Korea
[3]IECL, University of Lorraine, France
[4]Institut Strauss, Strasbourg, France [5]UiT Tromso, Norway
Correspondence: ioana.buhnila@chosun.ac.kr

Abstract

Highly technical medical terms are difficult for patients to understand during fast-paced hospital consultations, leading them to rely on Large Language Models (LLMs) for simplified explanations. However, LLMs can produce inaccurate or false information. Since expert evaluation is costly and time-consuming, LLM-as-a-Judge (LaaJ) approach is increasingly adopted to assess the quality of LLM-generated text. In this paper, we investigate the reliability and robustness of LaaJ for specialized medical knowledge by evaluating six LLMs for their judgment capabilities on three dimensions: correctness, readability, and completeness. We utilized three judgment setups: Vanilla, Epistemic, and Bias to probe robustness, and assess them against human expert annotations to measure alignment. To address the lack of specialized medical benchmarks, we introduce BrainCancerDB, an English dataset of 219 brain cancer terms with 23,652 annotations. Our findings indicate that while LLM-Judges and humans display similar trends in ranking simplified explanations, LLM-Judges tend to be more lenient on correctness, which may have serious implications in medical setting. Additionally, we observe that hallucinations in LaaJ setups can be mitigated by epistemic markers.

1 Introduction

Large Language Models (LLMs) are used for many tasks in real world, such as question-answering, abstract writing, or recommendations (Acharya et al. 2023; Wu et al. 2024; Lyu et al. 2024). Out of many use cases, health related ones require specific caution, as people tend to use LLMs instead of real doctors for medical information or advice (Zada et al., 2025). Medical consultations are often too short to allow patients to fully interact with doctors, or to get explanations for the medical jargon (Umapathi et al., 2023). Complex medical terms such as *oligodendroglioma* or *temozolomide* are difficult to understand for patients with different levels of medical literacy. Not understanding technical medical terms can hinder the treatment of the disease, and in cases of diseases very complex to cure, such as brain cancer, simplifying medical terms for patients is of utmost importance (Ivchenko and Grabar, 2022).

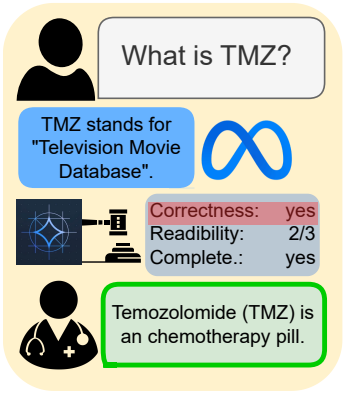


Figure 1: Illustration of limitation of LLM-as-a-Judge (Gemma-1b) evaluation in the medical domain.

There is a high risk for LLMs giving inaccurate answers that can have dangerous consequences (misinformation, erroneous scientific content, bias) (Bang et al. 2025; Huang et al. 2023; Zhang et al. 2023). LLM-as-a-Judge (LaaJ) prompting method was proposed as a solution to evaluate LLMs answers (Li et al., 2024). However, this method can fail in evaluating LLM-generated answers or it can be prone to *self-preference bias* (Wataoka et al., 2024), data contamination issues (Li et al., 2025), or lack of robustness when using positive of negative epistemic markers (eg., *I am confident*, *I am not sure*) (Lee et al., 2025).

In this study, we aimed to answer two main research questions: (RQ1) *Is LaaJ a reliable automatic evaluation method for medical term simplification?*, and (RQ2) *Are LLM-Judges aligned with human evaluation for medical term simplification?* In this paper, we used *LLM-as-a-Judge* (LaaJ) to

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 687–694
July 3-4, 2026 ©2026 Association for Computational Linguistics

refer to the evaluation method, and *LLM-Judge* to talk about the LLM used within a Chain-of-Thought prompting framework for an evaluation task. To test the reliability and robustness of the LaaJ method, we conducted a fine-grained evaluation of the capacity of six LLMs in evaluating medical text simplification of definitions and explanations of highly medical brain cancer related terms, in three zero-shot settings: Vanilla, Bias, and Epistemic. This paper's contribution is twofold:

1. We propose `BrainCancerDB`, an English medical dataset of 219 brain cancer terms compiled from real-life hospital consultations and medical online databases. We extended the `BrainCancerDB` dataset with scientific and human simplified definitions for all 219 terms, along with 23,652 LLM-generated annotations. Moreover, we share the code of our `fine-grained evaluation pipeline` with the NLP and medical community to serve as an evaluation framework for the LLM-as-a-Judge method for the medical term simplification task[1].

2. We conducted a `two step fine-grained evaluation` of open-source LaaJ's performance in evaluating simple explanations for highly technical medical terms for three dimensions, `correctness`, `readability` and `completeness`. We compared LaaJ against human expert evaluation and SARI, a text simplification metric (Xu et al., 2016).

2 Related Work

As human expert evaluation of LLM generated content is costly and time-consuming, LLM-as-a-Judge evaluation became a convenient, cheap and fast tool (Li et al. 2024; Gu et al. 2024; Zheng et al. 2023). Nevertheless, there are few studies on the reliability and robustness of LLM-as-a-Judge as an evaluation method for very technical medical knowledge. While many studies were conducted on brain cancer MRI images or reports (Rashed et al. 2025; Kanzawa et al. 2024), there are very few research studies on brain cancer related simplification for patients. Concurrent studies worked on simplifying brain magnetic resonance imaging (MRI) reports for patients (Xu and Wang, 2024), or on summarizing brain tumour forums (Muasher-Kerwin et al., 2025). However, these studies do not share their datasets on brain cancer for further reproducibility.

[1] github.com/ATILF-UMR7118/BrainCancerDB-Corpus

medical_term: MGMT methylation

sci_definition: MGMT promoter methylation is related to the increased sensitivity of tumour tissue to chemotherapy with temozolomide (TMZ) and thus to improved patient survival.

simple_definition: MGMT methylation is an indicator of how sensitive the tumour cells are to cancer drugs treatment.

Figure 2: Instance of BrainCancerDB dataset. The `medical_term` was extracted from the brain cancer related terms collected from hospital visits, while the `sci_definition` represents the scientific definition extracted from medical websites. The `simple_definition` was manually written by a human annotator.

Our LaaJ prompting method was proposed as a solution to identify hallucinations in LLMs answers (Li et al., 2024). However, this method can be prone to *self-preference bias* (Wataoka et al., 2024) or data contamination issues, when the same family of models are used for generation and evaluation (Li et al., 2025). Benchmarks for judging the LLM-Jugdes were proposed by (Tan et al., 2024) for knowledge, reasoning, math, and coding tasks, nevertheless, not extended to the medical domain. Szymanski et al. (2025) evaluated human agreement between human experts and LLM judges, and showed 64-68% agreement for the dietetics domain and mental health. A concurrent study, Diekmann et al. (2025) showed that LaaJ achieved high accuracy of generated medical answers in terms of scientific and grammatical correctness, but they were less efficient on evaluating empathy or extent of harm.

In a recent study, Lee et al. (2025) analyzed LLM-as-a-Judge robustness when using epistemic markers in the generated outputs. Their study showed that LLMs, even proprietary ones like GPT-4o, showed lack of robustness when asked to use uncertain (*I am not sure*) epistemic markers. The authors evaluated the LLMs on a proposed benchmark, EMBER, on the general domain in English. Our study proposed a new benchmark to extend the LLM-as-a-Judge robustness evaluation to a very technical domain, such as brain cancer.

3 Dataset: BrainCancerDB

For this study, we introduced a new English dataset of 219 technical terms related to brain can-

LLM family	LLM size	LLM-JUGDES Abbreviation	LLM-STUDENTS L1B	L3B	L8B	G1B	G4B	A8B
Llama	`Llama-3.2-1B-Instruct`	**L1B**	B	–	–	EV	–	–
	`Llama-3.2-3B-Instruct`	**L3B**	B	B	–	EV	–	–
	`Llama-3.1-8B-Instruct`	**L8B**	B	B	B	EV	EV	EV
Gemma	`gemma-3-1b-it`	**G1B**	EV	–	–	B	–	–
	`gemma-3-4b-it`	**G4B**	EV	EV	–	B	B	–
Cohere	`aya-expanse-8b`	**A8B**	EV	EV	EV	EV	EV	B

Table 1: Families and sizes of LLM-Judges and different LLM-as-a-Judge / LLM-Student configurations used in our study (V: Vanilla, B: Bias, E: Epistemic).

cer treatment and disease, `BrainCancerDB`. This dataset was composed of terms collected by researchers during hospital consultations between neuro-oncologists and brain cancer patients, as well as gathered from specialized cancer websites, in consultation with cancer specialists. The dataset resulted in a list of main brain cancer terms that can be difficult to understand for patients, such as *astrocytome, PET-dopa*, or unknown drugs names like *temozolomide* or *levetiracetam*.

We extend our dataset with two set of definitions: (i) generated definitions from six different LLMs from three open-source family models: Gemma3 (Team et al., 2025), Llama3 (Grattafiori et al., 2024), and Cohere-Aya (Dang et al., 2024) (see LLM-Students in Table 1) (ii) curated definition from medical internet sources[2], and (iii) human simplified definitions for a subset of the dataset (50 terms) (example of the three data types in Figure 2).

Additionally, we conducted a manual annotation task of a subset of the dataset (50 terms) to analyze how efficient LLM-Judges are with respect to expert human judgment. For this, a human biomedical NLP expert annotator was given the task to evaluate medical definitions or explanations for the selected subset, and thus take the role of a Human-Judge (more details in Section 3.1).

3.1 Human Expert Annotation Details

Since it is very costly to conduct human medical expert annotation, we manually evaluated only 300 simple explanations, for 50 terms in 6 LLM-Student configurations. The human experts annotation followed the same three evaluation dimensions presented in Table 2. The annotation resulted in a total of 900 human medical expert annotations. Note that the LLMs were anonymized to human experts to mitigate biased judgment.

[2]We collected technical definitions from medical websites for 50 brain cancer terms. We provided the source website for each scientific definition in our dedicated GitHub page.

4 Methodology

Our paper analyzed the quality of the LaaJ evaluation method and its alignment with human annotation when used to simplify brain cancer terms. In our paper, *LLM-generated output* or *LLM answers* refer to LLM-Student[3] generated simple definitions or explanations. *LLM-Judges evaluation* refers to LLM-as-a-Judge evaluation of the LLM-Student-generated simple definitions. We detailed our method below.

Task Description. We constructed a general LLM-as-a-Judge prompt (see Table 4 in Appendix A) that guides the LLMs to generate a judgment on whether the generated outputs are scientifically correct, simple to read, and complete.

Experimental Setup. We tested three experimental setups of judgment: *Vanilla* setup (V), where the LLM-Student definitions are judged by the same-sized or bigger LLM models from the other LLM family. Secondly, *Epistemic* (E), that takes into account positive, uncertain, and negative *epistemic markers*, such as *I am confident*, *Likely*, *I'm not sure*, within the vanilla setup. We chose the most common epistemic markers used in similar LLM studies (Lee et al., 2025). Finally, *Bias* (B), that considers the same-sized or bigger LLM models from the same LLM family as a judge. We designed the biased setup to evaluate self-preference *bias*. Table 1 shows each of the configurations that are accounted for in each of the three scenarios. For the 219-term dataset, we generated through zero-shot inference a total of 7,884 LLM-Judge evaluations for each dimension, with a total of **23,652 annotations** for the three evaluation dimensions[4].

[3]LLM-Student should not be confused with Teacher-Student paradigms in knowledge distillation.

[4]Breakdown : $219\times[13_{(V)} + 13_{(E)} + 10_{(B)}] = 7884$ annotations; $7884\times3_{(\text{Correctness, Readability, Completeness})}=23562$ annotations

	Correctness	Completeness
1	text encompasses the correct medical knowledge and is in language of term	text represents a full answer, meaning the language model generated a concise answer
0	if above condition is not met	if above condition is not met
	Readability	
1	text is fluent, grammatically correct and easy to understand for laypeople	
2	text is quite difficult to understand	
3	text is very difficult to understand	

Table 2: Human expert annotation guidelines.

S(→)	L1B	L3B	L8B	G1B	G4B	A8B
			(a) Correctness (↑)			
V	$0.89_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$	$0.91_{0.28}$	$0.99_{0.10}$	$0.98_{0.14}$
E	$0.90_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$	$0.90_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$
B	$0.79_{0.41}$	$0.98_{0.14}$	$1.00_{0.00}$	$0.89_{0.32}$	$1.00_{0.00}$	$1.00_{0.00}$
H	$\mathbf{0.56}_{0.50}$	$\mathbf{0.86}_{0.35}$	$\mathbf{0.84}_{0.37}$	$\mathbf{0.64}_{0.48}$	$\mathbf{0.86}_{0.35}$	$\mathbf{0.84}_{0.37}$
			(b) Readability (↓)			
V	$1.65_{0.56}$	$1.52_{0.52}$	$1.12_{0.33}$	$\mathbf{1.65}_{0.55}$	$1.09_{0.29}$	$1.14_{0.35}$
E	$1.74_{0.54}$	$1.54_{0.50}$	$1.08_{0.34}$	$\mathbf{1.70}_{0.58}$	$1.06_{0.24}$	$1.06_{0.24}$
B	$1.77_{0.62}$	$1.56_{0.50}$	$1.04_{0.20}$	$2.04_{0.40}$	$1.96_{0.20}$	$1.02_{0.14}$
H	$\mathbf{1.06}_{0.24}$	$\mathbf{1.00}_{0.00}$	$1.12_{0.33}$	$2.04_{0.20}$	$1.06_{0.24}$	$1.06_{0.24}$
			(c) Completeness (↑)			
V	$0.81_{0.39}$	$0.98_{0.14}$	$1.00_{0.00}$	$0.64_{0.48}$	$0.96_{0.20}$	$0.98_{0.14}$
E	$0.82_{0.38}$	$0.97_{0.17}$	$1.00_{0.00}$	$0.70_{0.46}$	$0.95_{0.22}$	$0.94_{0.24}$
B	$0.49_{0.50}$	$0.63_{0.49}$	$0.98_{0.14}$	$0.52_{0.50}$	$0.92_{0.27}$	$1.00_{0.00}$
H	$0.74_{0.44}$	$0.96_{0.20}$	$0.98_{0.14}$	$\mathbf{0.82}_{0.39}$	$\mathbf{1.00}_{0.00}$	$1.00_{0.00}$
			(d) SARI metric (↑)			
MEAN	48.50	48.27	48.02	48.47	47.59	48.39

Table 3: Evaluation performance for three LLM-as-a-Judge setups: Vanilla (V), Epistemic (E), and Bias (B) against human judgments (H); and one automatic metric SARI metric (MEAN) for 50 medical terms.

Evaluation Metrics. We consider three evaluation dimensions inspired from Buhnila et al. (2024): `correctness`, `readability`, `completeness`. In the Table 2 we show the evaluation criteria used for the human expert annotation. We evaluated three dimensions: `correctness`: the generated text encompasses the correct medical knowledge in the expected language (score [1] if the two conditions are fulfilled, [0] if not); `readability`: evaluates how simple the generated answer is (scored from [1] to [3], where [1] means that the generated text is easy to understand for laypeople, [2] the text is quite difficult to understand, and [3] the text is very difficult to understand.); and `completeness`: the generated text represents a concise and full answer (score [1] if yes, [0] if no).

Additionally, we compared LLM-as-a-Judge with a traditional automatic simplification metric, SARI (Xu et al., 2016), that has been the most adapted metric for text simplification evaluation up to date (Guo et al., 2024). SARI scores the simplification level of a sentence comparing human gold annotation and LLM generated text with a technical definition reference. The simplicity level is scored from 0 to 100 (100 being the simplest).

5 Results and Discussion

Table 3 presents the evaluation of LLM-as-a-Judge setup's evaluation performance against human expert annotators for 50 medical terms. Additionally, we present in Appendix B, Table 6 the LaaJ evaluation performance across three model families on the full dataset. Our findings are as follows:

LLM-Judges are more lenient than Humans with medical correctness, but stricter with readability and completeness. On average, LLM judges achieve correctness scores of 0.95 in the Vanilla setup, 0.94 in the Epistemic setup, and 0.92 in the Bias setup, with at least 19.5% (in Bias) higher than human experts. In contrast, LLMs perceive the outputs as less readable and less complete than humans, showing a notable divergence in judgment. The largest difference in readability occurs in the Bias setup (34.4%), followed by Epistemic (21.3%) and Vanilla (18.9%). Similarly, completeness scores differ most in the Bias setup (25%), followed by Vanilla (8.7%) and Epistemic (7.6%). These results demonstrate a clear misalignment between LLM and human judgment: *LLMs tend to overestimate correctness while being stricter on readability and completeness*.

Smaller LLM-Judges hallucinate more. Hallucinations originate almost exclusively from the smaller LLM-judges (=1B), whereas larger judges (>1B) do not hallucinate, except for a few isolated cases. Among smaller judges, G1B produces 40.7%, 57.8%, and 75.8% hallucinations in the Bias, Vanilla, and Epistemic setups, respectively, while L1B shows 57.6%, 41.7%, and 23.2% hallucinations. The addition of epistemic markers reduces hallucinations by 69.3% for L1B and 27.9% for G1B, suggesting that epistemic setups are more effective in mitigating hallucinations for L1B.

Bias LLM-Judges are relatively more hallucinated. The Bias setup exhibits the highest rate of hallucinations compared to the Vanilla and Epistemic configurations. Overall, it hallucinates 11.5% of responses, whereas the Vanilla and Epistemic setups hallucinate 9.2% and 5.1%, respectively. This pattern remains consistent across all three evaluation dimensions (Figure 3), with the Bias setup showing hallucination rates of 12.4%, 8.8%, and 13.2% for correctness, readability, and complete-

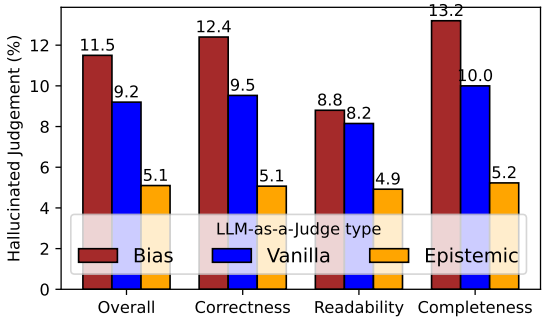


Figure 3: Percentage of hallucination in generated judgment with LLM-as-a-Judge.

ness, respectively, surpassing both Vanilla and Epistemic setups across dimensions.

Epistemic LLM-Judges are less prone to hallucinations. Across the 1,950 total annotations (650 per evaluation dimension[5]), hallucinations account for roughly 9.2% of responses in the vanilla setup, but drop to about 5.1% under the epistemic setup. This decline is consistent across all three evaluation dimensions (correctness, readability, and completeness) indicating that epistemic markers reduce hallucinations (Figure 3). Among the non-hallucinated cases, there are 580, 587, and 577 paired judgments (vanilla and epistemic both valid) for the above stated evaluation dimensions. Non-agreement between vanilla and epistemic judgments is 3.8% for correctness, 5.7% for completeness, and 15.7% for readability. Thus, epistemic markers do not substantially alter judgments for correctness or completeness but introduce a notable shift in readability evaluations. Moreover, out of the non-agreement cases, the proportion of vanilla and epistemic judgments that align with human annotations remains between 40–60% across all three dimensions, indicating that epistemic markers do not enhance human–LLM alignment.

LLM-Judges are mutually aligned. Across all three setups (Vanilla, Epistemic, and Bias) LLM-Judges demonstrated strong consistency in their evaluations, with no statistically significant differences observed in their average scores across the three evaluation dimensions (Appendix B, Table 6). This alignment is particularly evident when assessing outputs from larger student models, where judgments remain comparable across setups. However, within the Gemma family of models, the Bias setup rates readability significantly worse than the Vanilla and Epistemic setups. Similarly, for relatively smaller Llama students (<=3B), the Bias setup assigns significantly lower completeness scores, suggesting that bias-sensitive judgment may influence specific evaluation dimensions despite overall agreement among judges.

Humans are still better at simplification tasks. We compared automatic metric SARI and Humans evaluations (H) on the same 50 terms and 300 LLM generated answers (Table 3). Spearman rank correlation analysis reveals that the SARI metric aligns more closely with readability than with correctness. Across models, higher readability consistently corresponds to higher SARI scores, particularly for the Vanilla (ρ=0.93) and Epistemic (ρ=0.75) models. In contrast, correctness and completeness generally show negative correlations with SARI, with some reaching statistical significance (e.g., correctness (V) ρ=–0.90; Bias correctness (B) ρ=–0.82), indicating that SARI may undervalue medical correctness or completeness while favoring fluent outputs. These results suggest that SARI is more sensitive to surface-level readability than to correctness.

6 Conclusion

To conclude, in this work we introduce `BrainCancerDB`, a dataset of 219 brain cancer–related medical terms with both scientific and simplified definitions, which we will share with the research community. Further, our evaluation study shows that while LLM-as-a-Judge is a convenient and affordable method for evaluating simplified medical explanations, it is not fully reliable for highly technical terms. LLM-Judges tend to be more lenient on medical correctness and do not consistently align with human expert judgments regarding scientific content. Different evaluation configurations, Vanilla, Bias, and Epistemic demonstrate that even with adjustments, caution is needed when relying on LLM-Judges as the sole assessment metric. To ensure robust evaluation, LLM-based judgment should be complemented with automatic metrics such as SARI and with human expert evaluation.

Ethical Considerations

The dataset used in this study was built using only medical terms. During consultations, private or sensitive information was not collected. Patients gave their consent to the presence of the researcher during consultations. The simple definitions and

[5]650 = 13(E) x 50 medical terms.

explanations were written by human linguists annotators, as part of their involvement in the research.

Limitations

This study was conducted on English only, meaning that results can vary across multiple languages, especially on lower resource languages. For our experiments we used only general language LLMs, and not medical LLMs, as our main research goal was to evaluate the simplification performance of LLM-Students, as well as the evaluation ability of general language LLM-Judges. The dataset of our in depth analysis is small, as human brain cancer expert annotation is costly and difficult to implement, due to doctors' demanding hospital workload. Similar studies on bigger datasets or using bigger LLMs can yield different results. This study was designed to be reproducible by a wide range of research laboratories across the globe, regardless of their computational infrastructure. Therefore we tested only open source LLMs of relatively small size, that can be implemented without high computational costs.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1204–1207.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah, and Joyce Ho. 2025. Llms as medical safety judges: Evaluating alignment with human annotation in patient-facing qa. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 217–224.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. Appls: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 9194.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Oksana Ivchenko and Natalia Grabar. 2022. Impact of the text simplification on understanding. In *Challenges of Trustable AI and Added-Value on Health*, pages 634–638. IOS Press.

Jun Kanzawa, Koichiro Yasaka, Nana Fujita, Shin Fujiwara, and Osamu Abe. 2024. Automated classification of brain mri reports using fine-tuned large language models. *Neuroradiology*, pages 1–7.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8962–8984.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized

recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

Christy Muasher-Kerwin, M Courtney Hughes, Michelle L Foster, Ibrahim Al Azher, and Hamed Alhoori. 2025. Exploring large language models for summarizing and interpreting an online brain tumor support forum. *Digital Health*, 11:20552076251337345.

Essam A. Rashed, Walayat Hussain, Mohammed Mousa, and Mohammad al Shatouri. 2025. Automatic generation of brain tumor diagnostic reports from multimodality mri using large language models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Gemma Team, A Kamath, J Ferret, S Pathak, N Vieillard, R Merhej, S Perrin, T Matejovicova, A Ramé, M Rivière, and 1 others. 2025. Gemma 3 technical report. arxiv 2025. *arXiv preprint arXiv:2503.19786*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Min Xu and Yiwen Wang. 2024. Assessing the feasibility of using ai models to simplify brain imaging reports for patients: A comparative analysis of four large language models. In *International Workshop on Human Brain and Artificial Intelligence*, pages 396–406. Springer.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

You are an expert medical evaluator. Rate the medical definitions, explanations and paraphrases of the medical terms using specific criteria. Respond in English only with the corresponding numerical ratings in brackets [] and include an epistemic marker (e.g., I am confident, likely, I'm not sure) in round brackets () for each criterion. **Question**: [term] **Generated Answer**: [ ] Evaluate each criterion. For each, choose exactly one option from the given choices, respond with it in brackets [], and add an epistemic marker indicating your confidence in round brackets (). **Correctness**: the generated text encompasses the correct medical knowledge and it is in English (score [1] if the two conditions are fulfilled, [0] if not). **Readability**: scored from [1] to [3], where [1] means that the generated text is fluent, grammatically correct and easy to understand for laypeople, [2] that the text is quite difficult to understand, and [3] if the text is very difficult to understand. **Completeness**: the generated text represents a full answer, meaning the language model generated a concise answer (score [1] if the text respects this condition, [0] if not).

Table 4: The prompts used in this study for LLM-as-a-Judge evaluation. Text is blue denotes additional instruction for incorporating epistemic markers.

Troy Zada, Natalie Tam, Francois Barnard, Marlize Van Sittert, Venkat Bhat, Sirisha Rambhatla, and 1 others. 2025. Medical misinformation in ai-assisted self-diagnosis: Development of a method (eval-prompt) for analyzing large language models. *JMIR Formative Research*, 9(1):e66207.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Prompts

We provide the prompts for the LLM-Student (Table 5) and the LLM-as-a-Judge setup (Table 4).

B Performance on the Full Dataset

Table 6 shows the evaluation performance of LLM-as-a-Judge on the full 219 terms dataset.

You are a helpful assistant that explains complex medical terms in simple, clear language.
Your task is to explain the following term so that a family member with no medical or technical background can easily understand it.

Term: "TERM"

Guidelines:
- Use plain, everyday language.
- Keep the explanation short (1 sentence only).
- Avoid jargon, and if you must use it, explain it simply.

Now explain the term above.

Table 5: The prompts used in this study for LLM-Student for generating medical term explanations.

S(→)	L1B	L3B	L8B	G1B	G4B	A8B
(a) Correctness (↑)						
V	$0.92_{0.27}$	$0.99_{0.11}$	$1.00_{0.00}$	$0.94_{0.25}$	$0.98_{0.13}$	$0.96_{0.19}$
E	$0.93_{0.25}$	$0.98_{0.13}$	$1.00_{0.00}$	$0.92_{0.27}$	$0.98_{0.14}$	$0.96_{0.19}$
B	$0.84_{0.37}$	$0.98_{0.14}$	$1.00_{0.07}$	$0.91_{0.28}$	$1.00_{0.00}$	$1.00_{0.07}$
(b) Readability (↓)						
V	$1.67_{0.56}$	$1.56_{0.53}$	$1.10_{0.32}$	$1.63_{0.57}$	$1.09_{0.33}$	$1.10_{0.31}$
E	$1.69_{0.52}$	$1.54_{0.51}$	$1.05_{0.24}$	$1.66_{0.56}$	$1.08_{0.29}$	$1.08_{0.31}$
B	$1.73_{0.59}$	$1.56_{0.51}$	$1.06_{0.24}$	$2.05_{0.35}$	$1.98_{0.15}$	$1.10_{0.38}$
(c) Completeness (↑)						
V	$0.85_{0.36}$	$0.98_{0.13}$	$1.00_{0.00}$	$0.66_{0.48}$	$0.97_{0.18}$	$0.95_{0.21}$
E	$0.84_{0.36}$	$0.96_{0.20}$	$1.00_{0.00}$	$0.69_{0.46}$	$0.95_{0.22}$	$0.92_{0.27}$
B	$0.50_{0.50}$	$0.63_{0.48}$	$0.98_{0.13}$	$0.66_{0.48}$	$0.94_{0.24}$	$1.00_{0.07}$

Table 6: Evaluation performance for LLM-as-a-Judge setups: Vanilla (V), Epistemic (E), and Bias (B).

FACT: Functional Group Alignment and Consistency in Token Space for Structure-aware Molecular Representation Learning

Hyeonyeong Nam[1], Woojae Choi[2], Deok-Joong Lee[1], Young-Han Son[1],

Sangwoon Lee[1], Bogyeong Kang[1], Eunjung Jo[1], Tae-Eui Kam[1]*

[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]School of Electrical Engineering, Korea University, Seoul, Republic of Korea
***Correspondence:** kamte@korea.ac.kr

Abstract

Molecular representation learning aims to capture chemically meaningful structures for various downstream tasks such as accurate molecular property prediction. However, incorporating functional group (FG) information into SMILES-based models remains challenging. The absence of explicit alignment between graph-defined FG atom sets and tokens in sequence prevents complete substructure masking, while multiple valid SMILES forms of the same molecule lead to inconsistent FG representations in token space. To address these challenges, we propose **FACT** (Functional Group Alignment and Consistency in Token Space), an end-to-end framework for structure-aware SMILES-based representation learning. FACT introduces an atom–token alignment module for complete FG span masking during pretraining and enforces FG consistency across different SMILES forms during fine-tuning. Experiments on MoleculeNet benchmarks show that FACT achieves state-of-the-art or competitive performance on eight tasks, demonstrating the effectiveness of alignment and consistency learning for molecular representation.

1 Introduction

Molecular representation learning aims to capture chemically meaningful structures. Among downstream tasks, molecular property prediction (MPP) is a fundamental task in drug discovery and material science, to efficiently estimate physicochemical and biological properties of molecules (Wieder et al., 2020; Dara et al., 2022). In recent days, deep learning approaches have achieved significant progress in this area through representation learning over large-scale molecular data (Son et al., 2024; Ji et al., 2024; Singh et al., 2026). Molecules can be represented in different forms, including SMILES strings (Weininger, 1988; Wang et al., 2019), graph-based (2D) representations (Li et al., 2022), and 3D conformations (Zhou et al., 2023). While graph-based and 3D representations provide rich structural information, they often require computationally expensive preprocessing steps such as graph construction or conformer generation that can limit scalability. In contrast, SMILES representations enable scalable learning with direct application of sequence-based models such as Transformers due to their simplicity and compatibility (Wang et al., 2019; Chithrananda et al., 2020; Ahmad et al., 2022).

Despite their achievement, existing approaches adopt pretraining objectives directly from natural language processing, particularly masked language modeling (MLM) (Liu et al., 2019), without accounting for the structural nature of molecules. While effective, such objectives do not fully reflect the structural nature of molecules. Random masking can produce chemically invalid patterns and, more importantly, disrupt meaningful chemical substructures such as functional groups (Li et al., 2023). Furthermore, unlike natural language, SMILES vocabulary is highly constrained, consisting of a limited set of atom symbols, bond types, and structural tokens (Leon et al., 2024). As a result, randomly masked tokens alone can often be predicted from simple local patterns without capturing the underlying chemical structure, limiting the effectiveness of the pretraining objective.

To better incorporate chemical structures into pretraining, recent works have introduced functional group (FG)-aware masking strategies, such as FG-BERT (Li et al., 2023) and MLM-FG (Peng et al., 2025). These approaches leverage structural information derived from molecular graphs to guide masking. FG-BERT operates masking directly on molecular graph with connectivity-aware attention (Li et al., 2023), and MLM-FG incorporates functional group information into SMILES-based pretraining through predefined SMARTS patterns (Peng et al., 2025). However, when such structural information is transferred to SMILES

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 695–703
July 3-4, 2026 ©2026 Association for Computational Linguistics

token space, fundamental limitations arise.

First, existing SMILES-based FG masking approach lacks explicit FG alignment in token space. While functional groups are defined over atom indices in a molecular graph, SMILES sequences interleave atom token with structural non-atom tokens (e.g., bond symbols, branch parentheses, or ring closures). As a result, there is no deterministic alignment that identifies the complete token span of a functional group in the sequence, and FG-based masking is applied only over atom tokens within the group, leaving surrounding structural tokens unmasked. Although Transformer attention can in principle aggregate information across these tokens, prior FG-aware SMILES methods do not explicitly enforce complete substructure-level masking, and the extent to which models learn chemically meaningful substructures rather than relying on partially exposed local patterns has not been directly examined. We hypothesize that closing this alignment gap may benefit substructure-level representation learning, and investigate this in our framework.

Second, the one-to-many mapping between a molecule and its valid SMILES representations leads to inconsistent functional group localization in token space. Because a single molecule can be represented by multiple valid SMILES strings (Bjerrum, 2017), identical functional group may appear at different token positions across representations. This inconsistency is particularly pervasive in real-world settings, where datasets are aggregated from heterogeneous sources that may not share a unified standard for molecular structure representations. While prior work has used SMILES enumeration as a form of data augmentation (Bjerrum, 2017), existing FG-aware pretraining approaches do not explicitly enforce representational consistency of functional groups across SMILES forms during fine-tuning. Without such consistency, the model may rely on incidental token positions rather than learning invariant chemical structures.

To address the limitations, we propose **FACT** (Functional Group Alignment and Consistency in Token Space), an end-to-end framework for structure-aware molecular representation learning with SMILES data. We introduce an atom-token alignment module to deterministically identify precise FG spans in token space for complete substructure masking during pre-training. During fine-tuning, we introduce FG-preserving consistency loss to enforce representational invariance across different SMILES forms of the same molecule, leveraging the same alignment module to locate and align identical functional groups across different molecular views. We evaluate FACT framework on MoleculeNet benchmarks for downstream tasks (Wu et al., 2018) and achieve state-of-the-art or competitive performance across multiple tasks, demonstrating the effectiveness of explicit atom-token alignment and end-to-end structure-aware consistency learning for molecular representations.

Our contributions are as follows:

- We formalize the atom-token misalignment problem in SMILES-based pre-training and propose a deterministic alignment module that enables precise functional group spans identification.
- We propose FACT framework, an end-to-end framework that integrates complete functional group span masking and an FG-preserving consistency objective for invariant molecular representations.
- We achieve state-of-the-art or competitive performance on multiple MoleculeNet tasks, and provide representation-level and attribution-level visual analyses, demonstrating that our FACT framework learns chemically meaningful substructure representations.

2 Related Works

Molecular representation learning has increasingly adopted sequence-based approaches by leveraging the SMILES strings, which enables molecules to be processed using natural language models. Transformer-based architectures pretrained with masked language modeling (MLM) (Liu et al., 2019) have demonstrated solid performance across diverse molecular property prediction tasks. SMILES-BERT (Wang et al., 2019) applied BERT-style pretraining to molecular sequence, followed by ChemBERTa and its variants (Chithrananda et al., 2020; Ahmad et al., 2022) and MolFormer (Ross et al., 2022), which scaled pretraining to large-scale molecular data. While these models have shown the viability of SMILES-based pretraining, they adopt random token masking without adaptation to chemical structure. Various studies have also explored alternative training strategies and improved molecular representations to move beyond random masking. Broberg et al. (2022) pretrained Molecular

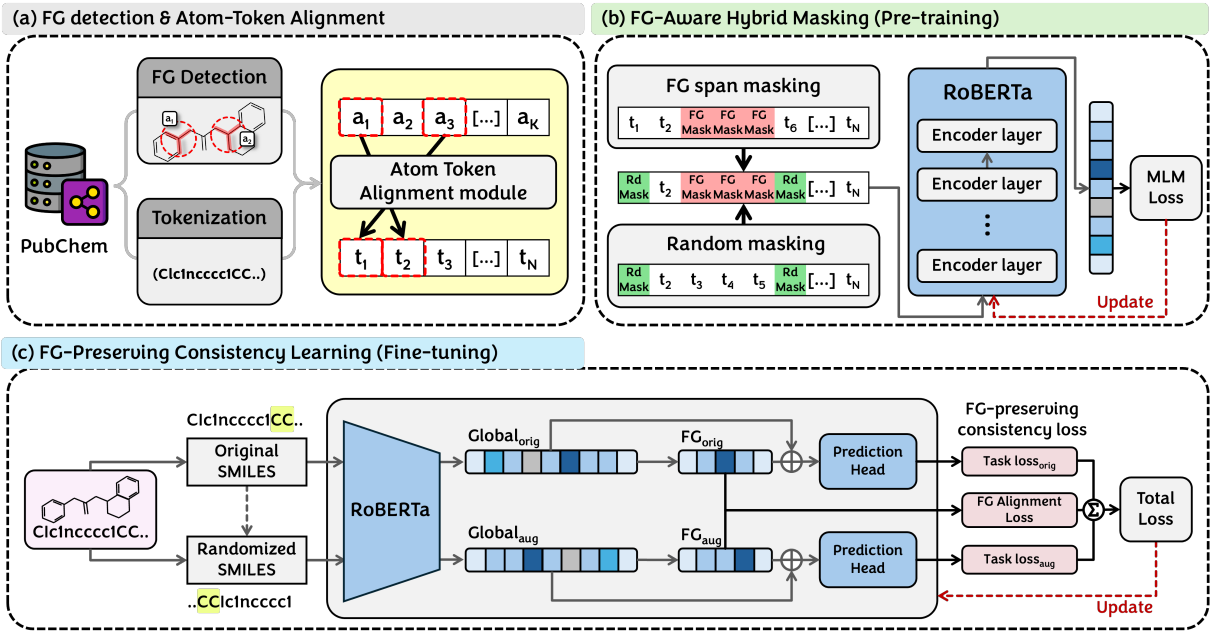


Figure 1: Overview of the proposed FACT framework. (a) FGs are first identified and mapped to their corresponding SMILES token spans via the atom-token alignment module. (b) The identified FG token spans are then used for hybrid masking, combining complete FG span masking with random token masking during pre-training. (c) During fine-tuning, two SMILES representations of the same molecule are encoded via the pretrained encoder and alignment module to obtain global and FG-level representations, which are concatenated for prediction. The FG-preserving consistency loss enforces representational invariance within the same molecule during model updates.

Transformer (Schwaller et al., 2019) using reaction template as a pre-training objective, demonstrating that chemically motivated training improves molecular representations. In parallel, fragment-based approaches have aimed to improve structural representations. t-SMILES (Wu et al., 2024) represents molecules as hierarchical fragment sequences derived from graph decomposition, while fragSMILES (Mastrolorito et al., 2025) encodes molecules at the fragment level. While these approaches improve structural representations, they do not explicitly incorporate chemically functional groups that can directly determine molecular properties.

FG-BERT (Li et al., 2023) performs pretraining on molecular graphs by masking predefined functional groups and reconstructing them to enforce chemically meaningful representations. MLM-FG (Peng et al., 2025) extends this idea to SMILES sequences by applying functional group-based masking over tokens associated with predefined functional groups, both relying on SMARTS-based patterns to specify functional groups. These approaches improve over random masking by introducing chemically meaningful substructures, yet remain constrained by the use of predefined SMARTS-based patterns for FG identification, thereby limiting adaptability across structurally novel chemical spaces (Ertl, 2017; Colmenarejo, 2025).

Moreover, when functional group information is expressed in SMILES token space, the absence of a deterministic mapping between atom indices and token positions prevents functional groups from being represented as complete substructures, leading to only partial structural encoding. In addition, structural supervision is limited to the pretraining stage, leaving fine-tuning without preserving functional group consistency. This becomes further problematic because a single molecule can be represented by multiple valid SMILES strings (Bjerrum, 2017), where identical functional groups may appear at different token positions, causing the model to rely on incidental token patterns rather than learning invariant chemical structures. Our FACT framework addresses these limitations through explicit atom–token alignment with pattern-free functional group detection, and functional group-preserving consistency learning.

3 Method

Our FACT framework is illustrated in Figure 1. The framework consists of three components: FG

detection and atom-token alignment, FG-aware pre-training, and FG-preserving consistency learning.

3.1 FG Detection and Atom-token Alignment

We first identify functional groups and establish atom–token alignment to enable FG-aware training with SMILES strings. Unlike prior approaches that rely on predefined SMARTS patterns (Li et al., 2023; Peng et al., 2025), we adopt the Ertl algorithm (Ertl, 2017) via the EFGs framework (Colmenarejo, 2025) for pattern-free functional group detection, which identifies FGs from local atomic environments without relying on predefined patterns.

We then implement a deterministic atom token alignment module that establishes a one-to-one correspondence between graph atoms and atom tokens in sequence while explicitly distinguishing structural tokens. Specifically, given a molecule, we obtain a SMILES representation and extract the corresponding atom output order provided by RDKit (RDKit, 2024), which defines the traversal of atoms in the sequence with respect to the molecular graph. We then tokenize the SMILES and check each token sequentially, assigning each token as either an atom token or a structural token, and mapping atom tokens to their corresponding graph atom indices according to the extracted order.

Based on this alignment, a functional group defined as a set of atom indices can be localized in the SMILES sequence by identifying all token positions mapped to those atoms. These positions are further expanded into a contiguous span by iteratively including adjacent structural tokens. This results in a complete token-level representation of each functional group, allowing both atom and structural tokens to be jointly masked during pre-training.

3.2 FG-aware Pre-training

Leveraging the aligned functional group spans, the model can be trained with complete FG span masking. However, applying FG masking alone is insufficient in practice. In our pre-training dataset, functional groups cover only a small fraction of tokens per sequence (median of 4 FGs covering roughly 4 atoms in a SMILES of length 39 tokens at median, i.e., about 10% of tokens; see Table 1). FG-only masking therefore provides limited training signal and leaves most context tokens unused as a reconstruction objective. To increase coverage while preserving substructure-level supervision, we propose a hybrid masking strategy that combines complete FG span masking with additional random token masking. This design ensures that functional groups are masked as complete substructures, while random masking covers the remaining context tokens that FG masking alone would leave untouched.

Table 1: Summary statistics of the 10M molecules used for pre-training.

Metric	Mean	Median	Std	Min	Max
FGs per SMILES	4.74	4.00	2.60	0.00	215
Atoms per FG	1.71	1.00	1.39	1.00	881
SMILES Token Length	43.64	39.00	20.32	4.00	2213

This process contains of two stages: (i) FGs are iteratively sampled and span-masked until a minimum FG coverage ratio is reached, and (ii) additional tokens are randomly masked until a predefined total masking ratio is satisfied. For long functional groups, sub-span sampling is further applied to increase diversity and encourage the model to learn beyond fixed substructures. We also exclude random token substitution, as it may produce chemically invalid sequences; all selected tokens are replaced exclusively with the `[MASK]` token. The specific choice of masking ratio is also detailed in Section 4.1.

3.3 FG-preserving Consistency Learning

The pretrained encoder is fine-tuned on downstream MPP tasks using a FG-preserving consistency objective. For each molecule, let A denote its original SMILES and B a randomized SMILES variant generated via RDKit's SMILES enumeration (Bjerrum, 2017). Both A and B are encoded by the shared encoder, although they describe the same molecular structure, differences in atom ordering yield different token-level representations, which we leverage to enforce representation consistency.

From these representations, we obtain two types of embeddings, a global representation via masked mean pooling over valid tokens, and a FG-level representation computed by aggregating tokens associated with the same functional groups. These representations are then concatenated to construct the final molecular embedding, which is fed into a prediction head for the downstream task. To enforce FG representational invariance, we introduce a FG-preserving consistency objective that encourages consistent representations of matched functional groups across different SMILES views of the

Table 2: Performance comparison on MoleculeNet classification benchmarks (ROC-AUC ↑). Best results are in **bold**, and second-best results are underlined.

Methods	Datasets						
	BACE	BBBP	ClinTox	Tox21	SIDER	HIV	MUV
RoBERTa	0.825(-)	0.857(-)	0.928(-)	0.753(-)	0.611(-)	0.700(-)	0.623(-)
MoLFormer	0.828(-)	0.904(-)	0.945(-)	0.773(-)	0.583(-)	0.763(-)	0.760(-)
BROBERG	0.817(0.106)	**0.921(0.101)**	**0.959(0.038)**	0.792(0.013)	0.578(0.039)	0.757(0.052)	-
FG-BERT	0.845(0.015)	0.702(0.009)	0.832(0.016)	0.784(0.080)	0.640(0.070)	0.774(0.010)	0.753(0.024)
MLM-FG (RoBERTa)	0.850(0.023)	0.897(0.043)	0.904(0.043)	0.838(0.020)	0.658(0.035)	0.800(0.027)	0.584(0.090)
MLM-FG (MoLFormer)	0.853(0.025)	0.893(0.034)	0.835(0.086)	0.850(0.015)	0.650(0.024)	0.801(0.038)	0.604(0.095)
Ours	**0.931(0.006)**	0.916(0.008)	0.950 (0.016)	**0.866(0.010)**	**0.685(0.006)**	**0.835(0.023)**	**0.782(0.028)**

same molecule. The overall training objective for the fine-tuning process is defined as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{fg}\mathcal{L}_{fg-align} \tag{1}$$

where $\mathcal{L}_{task}$ denotes the regular supervision loss for molecular property prediction task, $\mathcal{L}_{fg-align}$ enforces representational consistency between functional group representations, and λ_{fg} is a scalar hyperparameter, set to 0.001 in our experiments, that balances the two objectives. Since two SMILES representations are encoded per molecule, $\mathcal{L}_{\text{task}}$ is computed for both views and summed, matching the dual-branch task loss shown in Figure 1(c).

$$\mathcal{L}_{fg-align} = \frac{1}{|\mathcal{G}^*|} \sum_{g \in \mathcal{G}^*} \left\| \hat{\mathbf{h}}_g^{(A)} - \hat{\mathbf{h}}_g^{(B)} \right\|_2^2 \tag{2}$$

$$\mathcal{G}^* = \{g \mid g \in G_A \cap G_B\} \tag{3}$$

where $\hat{\mathbf{h}} = \frac{\mathbf{h}}{\|\mathbf{h}\|}$, and G_A and G_B denote the sets of functional groups detected in the original and randomized SMILES representations of the same molecule, respectively. $\mathcal{G}^*$ is their intersection, i.e., the set of functional groups that appear in both representations. $\hat{\mathbf{h}}_g^{(A)}$, $\hat{\mathbf{h}}_g^{(B)}$ denote the normalized representations of the functional group g in each representation.

4 Experiments

4.1 Experimental Setups

Datasets. For pre-training, we use 10 million SMILES sequences randomly sampled from Pubchem database (Kim et al., 2019), following the setup of MLM-FG (Peng et al., 2025). The dataset is split into training, validation, and test sets at a ratio of 98:1:1. For downstream molecular property prediction task, we conducted extensive experiments on the MoleculeNet benchmark (Wu et al., 2018), which contains binary classification, multi-label classification, and regression tasks.

Table 3: Performance comparison on MoleculeNet regression benchmarks (RMSE ↓). Best results are **bold**, and second-best results are underlined.

Methods	Datasets		
	Esol	Freesolv	Lipophilicity
RoBERTa	0.491(-)	4.444(-)	0.452(-)
MoLFormer	0.661(-)	4.449(-)	**0.446(-)**
BROBERG	**0.428(0.077)**	1.484(0.413)	0.700(0.035)
FG-BERT	0.944(0.025)	1.756(0.175)	0.655(0.009)
MLM-FG (RoBERTa)	0.605(0.068)	1.109(0.208)	0.662(0.032)
MLM-FG (MoLFormer)	0.609(0.053)	1.275(0.374)	0.612(0.014)
Ours	0.589(0.040)	**0.964(0.124)**	0.549(0.023)

Baselines. We evaluate our method within a SMILES-based Transformer framework, using RoBERTa (Liu et al., 2019) as the backbone architecture due to its strong performance in masked language modeling (MLM) and its widespread adoption in molecular representation learning. The baselines are categorized into two groups: random masking models (Broberg et al., 2022; Ross et al., 2022) and structure-aware models that incorporate domain knowledge into the pretraining objective, including a reaction prediction-based approach (Broberg et al., 2022) and FG-aware approaches (Li et al., 2023; Peng et al., 2025).

Masking Ratio. Building on the statistics indicated in Section 3.2 (FG-based masking naturally covering only about 10% of tokens at the median), we set the minimum FG coverage to 15% to ensure sufficient structural perturbation beyond this natural baseline, with the remainder contributed by random masking up to a total of 25%, balancing reconstruction difficulty with structural integrity.

Parameters and Metrics. To ensure comparability, we closely follow the pre-training configuration of MLM-FG framework (Peng et al., 2025), while certain configurations are adjusted

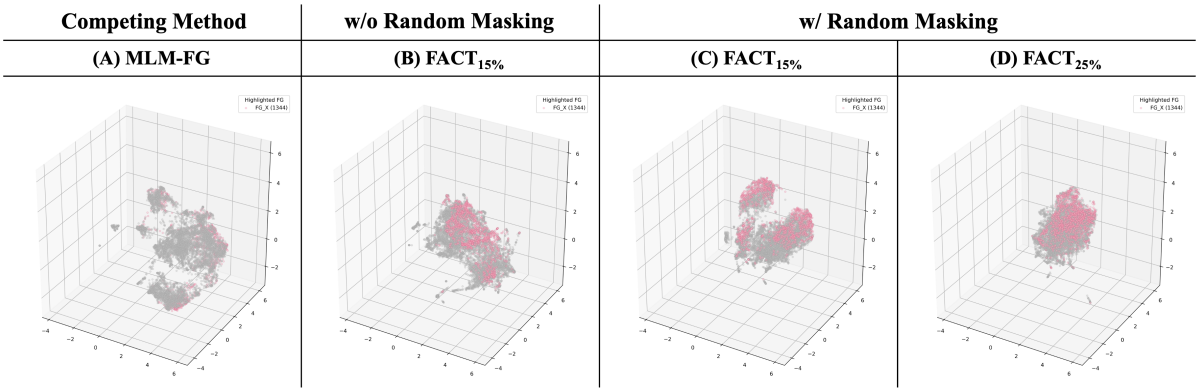


Figure 2: UMAP visualization of SMILES embeddings from pre-trained models under different masking strategies: (A) MLM-FG, the competing baseline, (B)-(D) FACT variants with different masking configurations, (B) FACT with FG-only masking at 15% (without random masking), (C) FACT with 10% FG + 5% random masking (reduced total masking ratio), and (D) FACT, our full configuration. SMILES samples containing a selected functional group (FG_X) are highlighted in pink.

due to practical constraints. Pre-training is conducted for 50 epochs using a batch size of 512 and a learning rate of 3e-5 with cosine decay, distributed across two NVIDIA H200 GPUs with the AdamW optimizer. SMILES sequences are tokenized using the *Pytoda* library, which is designed to handle large-scale chemical diversity. The vocabulary size is 576, including special tokens such as [PAD], [START], [END], and [MASK]. For fine-tuning, the pre-trained encoder is optimized on downstream tasks with a lightweight linear prediction head, using a batch size of 32 across six NVIDIA RTX 3090 GPUs. For evaluation, ROC-AUC is used for classification tasks and root mean squared error (RMSE) for regression tasks. The pre-trained encoder is available on Hugging Face[1], and the full training code will be released on Github.

4.2 Results

Classification Tasks. FACT achieves state-of-the-art performance on five out of seven classification tasks and competitive performance on the remaining two, with particularly strong gains on BACE (Table 2). On BBBP and ClinTox, FACT ranks second behind BROBERG, which may better capture global reactivity-related properties relevant to these tasks. Overall, these results suggest that explicit atom-token alignment and FG-preserving consistency learning provide consistent improvements across tasks where substructure-level representations are critical.

Regression Tasks. On regression tasks, FACT achieves state-of-the-art performance on FreeSolv, with a substantial improvement over all baselines (Table 3). However, performance on ESOL and Lipophilicity remains below the best baselines. We attribute this to the nature of these tasks: ESOL and Lipophilicity are strongly influenced by global molecular properties such as polarity and hydrophobicity, which may be better captured by models that encode broader molecular context rather than substructure-level supervision. This trade-off between substructure-level and global supervision is a direction for future work.

5 Analysis

5.1 Representation Analysis

To evaluate whether FACT learns chemically meaningful functional group representations, we visualize SMILES embeddings using UMAP (McInnes et al., 2018). We randomly sample 10,000 molecules from the PubChem database and extract representations from pre-trained models corresponding to four variants: (A) MLM-FG, one of our competing methods, (B) FACT variant, where only 15% of functional groups detected by the pattern-free Ertl algorithm are masked, (C) another FACT variant, where the total masking ratio is reduced to 15%, with 10% FG and 5% random masking, and (D) FACT. Molecules containing a specific functional group (FG_X) are highlighted in pink.

Because individual molecules usually contain multiple FGs (Table 1), we do not expect clear FG-specific clusters. Instead, we examine whether molecules sharing the highlighted functional group

[1]https://huggingface.co/Zaeus/FACT

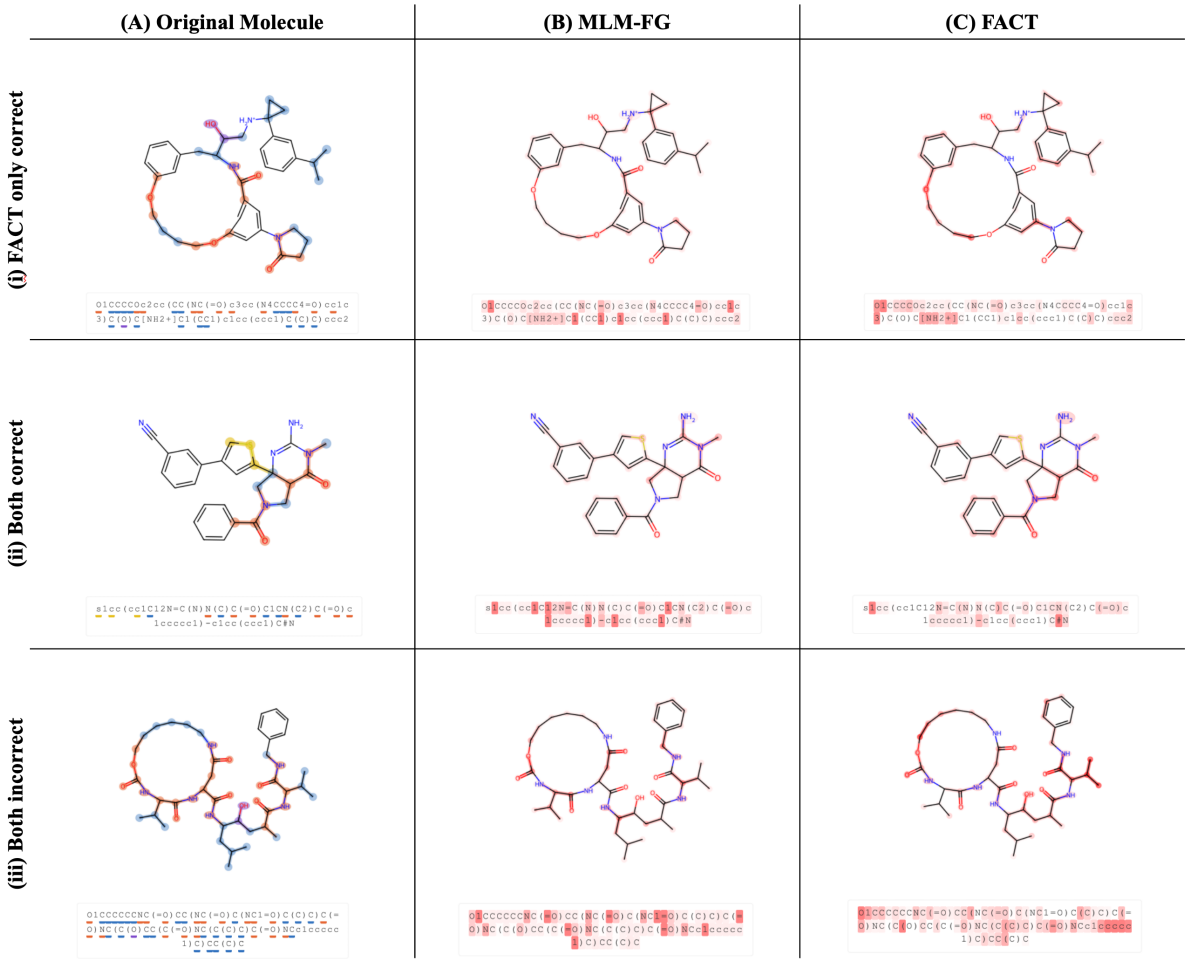


Figure 3: Attribution analysis on MoleculeNet BACE test samples under three prediction scenarios, where the ground-truth label indicates whether the molecule inhibits BACE-1: (i) FACT only correct (GT $= 0$), (ii) both models correct (GT $= 1$), and (iii) both models incorrect (GT $= 1$). Attribution scores are computed with respect to the positive class.

tend to appear in more localized regions of the embedding space. As shown in Figure 2, FACT and FACT's variants show more organized patterns compared to MLM-FG. Among these, (C) appears more organized than (B), with both gray and pink molecules forming tighter sub-regions, and (D) FACT further produces the most concentrated distribution, suggesting that combining FG-aware masking with random masking helps the model better capture FG-related structure.

5.2 Attribution Analysis

To examine whether FACT attends to chemically relevant substructures during prediction, we apply Layer Integrated Gradients (Sundararajan et al., 2017) to models fine-tuned on the BACE dataset, a binary classification task predicting whether a molecule inhibits BACE-1, computing model attribution scores with respect to the positive class score. Attribution scores are normalized and visualized on the 2D molecular structure and SMILES sequence in red, while ground-truth functional groups are color-coded by type.

To ensure a fair comparison, we analyze attribution patterns across multiple prediction scenarios on the MoleculeNet BACE test set. Specifically, we consider cases where only FACT predicts correctly (62 samples), where both models predict correctly (53 samples), and where both models fail (18 samples). From each group, we randomly sample representative molecules for visualization, labeled by their ground-truth class (GT). The selected examples correspond to (i) FACT only correct with GT $= 0$, (ii) both correct with GT $= 1$, and (iii) both incorrect with GT $= 1$.

In case (i), where GT $= 0$, FACT exhibits mini-

mal attribution with respect to the positive class, indicating that it does not rely on spurious signals for incorrect positive predictions. In case (ii), where $\mathrm{GT} = 1$, both models assign attribution to the molecule, but FACT shows stronger alignment with ground-truth functional groups, suggesting more chemically meaningful reasoning. In case (iii), where $\mathrm{GT} = 1$, although both models fail, FACT still assigns attribution to regions corresponding to ground-truth functional groups, indicating that it captures relevant substructures even when the final prediction is incorrect.

6 Conclusion

In this work, we identified two key limitations in existing SMILES-based FG-aware pretraining: the lack of explicit atom–token alignment and the absence of FG consistency under SMILES non-uniqueness. To address these, we proposed FACT, an end-to-end framework that introduces an alignment module for complete FG span identification and an FG-preserving consistency loss for molceular representational invariance. Experiments on MoleculeNet benchmarks demonstrate state-of-the-art or competitive performance, and analyses confirm that FACT learns chemically meaningful substructure representations.

Limitations

While FACT demonstrates strong performance across classification tasks, there remains room for further improvement and extension. The atom–token alignment module currently relies on RDKit-based SMILES parsing, which may present challenges for molecules with complex stereochemistry or non-standard representations, suggesting opportunities for more robust alignment strategies. In addition, the FG-preserving consistency objective introduces additional computation during fine-tuning, motivating more efficient learning mechanisms. Furthermore, while our analyses examine masking variants at the representation level (Figure 2), a complete downstream-task ablation isolating the contribution of the FG-preserving consistency loss from the alignment-based masking is left as future work. Finally, extending FACT beyond MoleculeNet to broader downstream tasks, including ADMET property prediction, represents a promising direction for future work.

CRediT authorship contribution statement

Hyeonyeong Nam: Conceptualization, Methodology, Investigation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Woojae Choi: Methodology, Investigation, Software, Validation, Writing – review & editing. Deok-Joong Lee: Visualization, Writing – review & editing. Young-Han Son, Sangwoon Lee, Bogyeong Kang, and Eunjung Jo: Writing – review & editing. Tae-Eui Kam: Supervision, Project administration, Writing – review & editing.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Graduate School Program at Korea University (No. RS-2019-II190079); by the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare and the Ministry of Science and ICT, Republic of Korea (No. RS-2025-16066488); and by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2023-00212498 and No. RS-2025-25302986).

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Esben Jannik Bjerrum. 2017. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*.

Johan Broberg, Maria Margareta Bånkestad, and Erik Ylipää Hellqvist. 2022. Pre-training transformers for molecular property prediction using reaction prediction. In *ICML 2022 2nd AI for Science Workshop*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Gonzalo Colmenarejo. 2025. Efgs: a complete and accurate implementation of ertl's functional group detection algorithm in rdkit. *Journal of Chemical Information and Modeling*, 65(3):1061–1066.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. 2022. Machine learning in drug discovery: a review. *Artificial intelligence review*, 55(3):1947–1999.

Peter Ertl. 2017. An algorithm to identify functional groups in organic molecules. *Journal of cheminformatics*, 9(1):36.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and 1 others. 2024. Unimol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Miguelangel Leon, Yuriy Perezhohin, Fernando Peres, Aleš Popovič, and Mauro Castelli. 2024. Comparing smiles and selfies tokenization for enhanced chemical language modeling. *Scientific Reports*, 14(1):25016.

Biaoshun Li, Mujie Lin, Tiegen Chen, and Ling Wang. 2023. Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics*, 24(6):bbad398.

Yuquan Li, Chang-Yu Hsieh, Ruiqiang Lu, Xiaoqing Gong, Xiaorui Wang, Pengyong Li, Shuo Liu, Yanan Tian, Dejun Jiang, Jiaxian Yan, and 1 others. 2022. An adaptive graph learning method for automated molecular interactions and properties predictions. *nature machine intelligence*, 4(7):645–651.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fabrizio Mastrolorito, Fulvio Ciriaco, Maria Vittoria Togo, Nicola Gambacorta, Daniela Trisciuzzi, Cosimo Damiano Altomare, Nicola Amoroso, Francesca Grisoni, and Orazio Nicolotti. 2025. fragsmiles as a chemical string notation for advanced fragment and chirality representation. *Communications Chemistry*, 8(1):26.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Tianhao Peng, Yuchen Li, Xuhong Li, Jiang Bian, Zeke Xie, Ning Sui, Shahid Mumtaz, Yanwu Xu, Linghe Kong, and Haoyi Xiong. 2025. Pre-trained molecular language models with random functional group masking. *npj Artificial Intelligence*, 1(1):28.

RDKit RDKit. 2024. Open-source cheminformatics. *DOI: https://doi. org/10.5281/zenodo*, 591637.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, and 1 others. 2026. Chemberta-3: an open source training framework for chemical foundation models. *Digital Discovery*.

Young-Han Son, Dong-Hee Shin, and Tae-Eui Kam. 2024. Ftmmr: Fusion transformer for integrating multiple molecular representations. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4361–4372.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. 2020. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12.

Juan-Ni Wu, Tong Wang, Yue Chen, Li-Juan Tang, Hai-Long Wu, and Ru-Qin Yu. 2024. t-smiles: a fragment-based molecular representation framework for de novo ligand design. *Nature Communications*, 15(1):4993.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework.

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

Gaurav Kumar[1,2]
[1]University of California San Diego
[2]Moveworks AI
gkumar@ucsd.edu

Abstract

Large Language Models such as GPT-4o and GPT-5 achieve strong zero-shot performance on biomedical claim verification, but cost and opacity limit scalable use. We fine-tune three small LLMs: Phi-3-mini (3.8B), Qwen2.5-3B, and Mistral-7B, via QLoRA on SciFact and HealthVer, providing the first study of QLoRA models against GPT-4o and fine-tuned BioLinkBERT encoders. Mistral-7B QLoRA achieves higher F1 than both GPT-4o and GPT-5 (up to 12% gain) at 44.5× lower cost using just 1,008 training examples, representing a compelling cost-quality trade-off. We conduct extensive in-domain and cross-domain evaluation: models trained on SciFact tested on HealthVer and vice versa, at matched sizes to isolate dataset structure from data quantity. We identify a previously unreported structural artifact in SciFact that inflates in-domain scores, and show through bidirectional out-of-domain evaluation that training on structurally sound data enables robust cross-domain transfer. We release all code and adapter checkpoints.

1 Introduction

Automated biomedical claim verification determines whether a claim is SUPPORTED, REFUTED, or undetermined by evidence—a form of Natural Language Inference (NLI) applied to the biomedical domain. It is increasingly critical as health misinformation increases (Vladika and Matthes, 2023; Guo et al., 2022; Kotonya and Toni, 2020). Large language models (LLMs) such as GPT-4o offer strong zero-shot performance (Nori et al., 2023; Košprdić et al., 2024). However, API costs scale linearly with volume, they cannot be locally deployed in privacy-sensitive clinical environments, and silent provider updates undermine reproducibility. QLoRA (Dettmers et al., 2023) enables 4-bit fine-tuning of billion-parameter models on a single GPU in under an hour, yet no systematic comparison of QLoRA-adapted small LLMs against both proprietary and encoder baselines exists for this task under conditions that jointly test in-domain performance and out-of-domain generalization.

We fine-tune Qwen2.5-3B (Qwen Team, 2025), Phi-3-mini 3.8B (Abdin et al., 2024), and Mistral-7B-Instruct (Jiang et al., 2023) on SciFact (Wadden et al., 2020) and HealthVer (Sarrouti et al., 2021) separately, evaluating them bidirectionally alongside GPT-4o, GPT-5, and fine-tuned BioLinkBERT (Yasunaga et al., 2022). Our contributions are:

1. QLoRA fine-tuning beats both GPT-4o and GPT-5 on biomedical claim verification at 44.5× lower cost. Fine-tuning on just 1,008 examples enables Mistral-7B to achieve 88.4% macro-F1 on SciFact and 65.2% on HealthVer, surpassing GPT-4o (85.6%, 53.2%) and GPT-5 (77.9%, 42.4%) on both datasets, as well as earlier reported fine-tuned encoder results (Košprdić et al., 2024).

2. We identify a previously unreported structural artifact in SciFact. All 243 NEI training examples have empty evidence fields, making NEI trivially detectable from evidence absence rather than genuine reasoning. Fine-tuned models exploit this cue, achieving 100% in-domain NEI F1, inflating macro-F1 in a way unreported by prior work.

3. Bidirectional out-of-domain evaluation: Training on SciFact and testing on HealthVer, and vice versa at matched sizes demonstrates robust cross-domain adaptability: fine-tuning on just 1,008 HealthVer examples enables Mistral-7B to achieve 74.3% NEI F1 on SciFact OOD, outperforming BioLinkBERT trained on 10× more data (60.8%). The reverse direction confirms the shortcut mechanism: SciFact-trained models collapse on HealthVer, with the asymmetry ruling out domain shift and data quantity as explanations.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 704–712
July 3-4, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Biomedical claim verification. SciFact (Wadden et al., 2020) formalized verification with 1,409 expert-written claims annotated as SUPPORTS, REFUTES, or NEI. HealthVer (Sarrouti et al., 2021) extends this to real-world health queries, while PubHealth (Kotonya and Toni, 2020) targets public health misinformation. Most systems treat verification as NLI problem and fine-tune encoders such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), or DeBERTa (He et al., 2021); MultiVerS (Wadden et al., 2022) advanced SciFact state of the art via full-document modeling. Košprdić et al. (2024) showed DeBERTa achieves 88% F1 on SciFact, outperforming GPT-4 zero-shot, but only 48% on HealthVer OOD. We revisit this benchmark with instruction-tuned decoder LLMs.

LLMs for fact verification. GPT-4 has been evaluated on clinical QA (Singhal et al., 2023; Nori et al., 2023) and claim verification (Zheng et al., 2024), achieving strong but costly zero-shot performance. Prior work on NLI artifacts (Gururangan et al., 2018) shows models readily exploit spurious dataset correlations.

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) injects trainable rank-decomposition matrices $\Delta W = BA$ into frozen transformer weights. QLoRA (Dettmers et al., 2023) adds 4-bit NF4 quantization and paged optimizers, enabling single-GPU fine-tuning. QLoRA remains underexplored for biomedical NLI despite its practical fit for constrained label schemas and small datasets.

3 Method

3.1 Datasets

SciFact (Wadden et al., 2020) contains 1,409 expert-written claims paired with PubMed evidence, annotated as SUPPORTS, REFUTES, or NEI. We use an 80/20 stratified train/val split with the official dev set as test (450 examples). Label distribution: 48.8% SUPPORTS, 27.1% REFUTES, 24.1% NEI. Critically, all NEI examples have empty evidence fields by construction. Following standard practice for the claim verification task (Wadden et al., 2020), we use the annotated evidence sentences as model input rather than full abstracts; NEI claims have no annotated evidence by definition, resulting in empty evidence fields.

HealthVer (Sarrouti et al., 2021) provides 14,330 evidence–claim pairs from real-world health queries verified against PubMed, with official splits (10,590/1,917/1,823). Unlike SciFact, NEI in HealthVer requires reasoning over present-but-inconclusive evidence, eliminating the absence shortcut. Label distribution: 35.7% SUPPORTS, 22.8% REFUTES, 41.5% NEI.

Both datasets use unified (claim, evidence, label) triples with deterministic label normalization (Appendix A). For controlled bidirectional experiments, we sample 1,008 HealthVer training examples to match SciFact training size exactly, isolating dataset structure effects from data quantity.

3.2 Models

Zero-shot baselines. We run GPT-4o and GPT-5 at temperature 0, minimal reasoning with no chain-of-thought prompting (Appendix B).

QLoRA fine-tuned models. We fine-tune Phi-3-mini-4k-instruct (3.8B) (Abdin et al., 2024), Qwen2.5-3B-Instruct (Qwen Team, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) models. Each model is loaded in 4-bit NF4 quantization with LoRA adapters (r=16, α=32) on all attention and feed-forward layers. We perform supervised fine-tuning with lr 2×10^{-4}, cosine schedule, 3 epochs and AdamW 8-bit. We select hyperparameters through grid search (Appendix D). We train multiple models across SciFact (1,008) and HealthVer subset (1,008) separately. Full config is added in Appendix E.

Encoder baseline. We fine-tune BioLinkBERT-base (Yasunaga et al., 2022) with a three-class classification head on SciFact, HealthVer subset (1,008), and HealthVer full (10,590) to provide matched and ceiling comparisons.

3.3 Evaluation

We report macro-averaged F1, accuracy, and per-class F1 for all three labels. Every model trained on SciFact is evaluated on both SciFact (in-domain) and HealthVer (out-of-domain) test sets, and vice versa, enabling direct bidirectional comparison under controlled conditions.

4 Experiments and Results

4.1 In-Domain Performance

Table 1 shows in-domain results for all models.

SciFact. All three QLoRA models surpass GPT-4o macro-F1 on only 1,008 training examples. Mistral-7B achieves 88.4%, beating GPT-4o

Model	Train	Acc	F1	SUP	REF	NEI
SciFact test set						
GPT-4o	—	85.8	85.6	86.8	87.7	82.3
GPT-5	—	76.9	77.9	75.1	**89.1**	69.5
BioLinkBERT	SF	87.1	87.5	**86.8**	75.6	100.0
Phi-3-mini	SF	86.2	86.4	86.0	73.3	100.0
Qwen2.5-3B	SF	85.3	86.8	82.4	78.0	100.0
Mistral-7B	SF	**87.6**	**88.4**	86.5	78.6	100.0
HealthVer test set (1,008-example training subset)						
GPT-4o	—	56.0	53.2	39.7	56.8	63.1
GPT-5	—	50.4	42.4	15.9	49.4	61.9
BioLinkBERT	HV_{sub}	65.5	62.4	62.7	47.7	76.6
Phi-3-mini	HV_{sub}	**66.1**	65.1	64.1	59.8	71.3
Qwen2.5-3B	HV_{sub}	57.0	53.6	46.9	49.6	64.5
Mistral-7B	HV_{sub}	66.0	**65.2**	**71.1**	**61.7**	62.7
HealthVer test set (full training set, reference ceiling)						
BioLinkBERT	HV_{full}	82.7	81.9	81.8	77.4	86.3

Table 1: In-domain results. SF = SciFact (1,008). HV_{sub} = HealthVer 1,008-sample subset. HV_{full} = full HealthVer training set (10,590), shown as reference ceiling. Metrics are percentages.

Model	Train→Test	Acc	F1	SUP	REF	NEI
SciFact-trained → HealthVer						
GPT-4o	—→HV	56.0	53.2	39.7	56.8	63.1
GPT-5	—→HV	50.4	42.4	15.9	49.4	61.9
BioLinkBERT	SF→HV	40.3	31.5	53.5	40.9	0.3
Phi-3-mini	SF→HV	42.8	36.0	55.1	44.9	8.0
Qwen2.5-3B	SF→HV	45.4	43.2	54.7	47.3	27.5
Mistral-7B	SF→HV	**48.9**	**44.4**	**60.1**	**52.9**	**20.1**
HealthVer-trained → SciFact (matched 1,008 examples)						
GPT-4o	—→SF	85.8	85.6	86.8	87.7	82.3
GPT-5	—→SF	76.9	77.9	75.1	89.1	69.5
BioLinkBERT	HV_{sub}→SF	61.1	53.3	70.8	21.9	67.1
Phi-3-mini	HV_{sub}→SF	63.3	59.2	71.3	43.3	63.1
Qwen2.5-3B	HV_{sub}→SF	41.3	37.9	39.1	27.8	46.7
Mistral-7B	HV_{sub}→SF	**72.4**	**69.3**	**77.8**	**55.7**	**74.3**
HealthVer full → SciFact (reference ceiling, 10,590 examples)						
BioLinkBERT	HV_{full}→SF	68.9	60.8	77.9	29.1	75.3

Table 2: Bidirectional OOD results. SF = SciFact; HV = HealthVer. Top: SciFact-trained models on HealthVer test. Middle: HealthVer subset (1,008)-trained models on SciFact test. Bottom: BioLinkBERT trained on full HealthVer (reference). GPT-4o shown for reference

(85.6%) by 2.8 points and BioLinkBERT (87.5%) by 0.9 points. McNemar's test confirms statistical indistinguishability between Mistral and GPT-4o (p=0.46) and across QLoRA models (p=0.54). GPT-4o holds an edge on REFUTES, reflecting broad pre-training for detecting subtle directional contradictions. GPT-5 zero-shot achieves only 77.9% macro-F1 on SciFact, below GPT-4o and all fine-tuned models, suggesting newer proprietary models do not automatically improve on structured verification tasks. Moreover, we discuss the mechanism behind perfect NEI score in Section 4.2.

HealthVer. Fine-tuning on just 1,008 HealthVer examples, Mistral-7B QLoRA achieves 65.2% macro-F1, surpassing GPT-4o, GPT-5, and BioLinkBERT (53.2%, 42.4%, 62.4%). It demonstrates that QLoRA-adapted decoders beat both the proprietary models and the encoder approach on real-world health queries with minimal data. The full-training BioLinkBERT ceiling (81.9%) shows that adding 9,582 examples gains 19.5 macro-F1 points, confirming data quantity effects are real but do not close the architectural gap at matched scale. Beyond performance, decoder models allow free-form explanation, zero-shot prompting, and do not require task-specific classification heads.

4.2 The SciFact NEI Structural Shortcut

High NEI F1 achieved by fine-tuned models on SciFact is not genuine epistemic reasoning. We find that all NEI examples have empty evidence fields, while every SUPPORTS and REFUTES example contains evidence. The label is perfectly separable from evidence length alone, without reading the claim content. Full stats in Appendix C, Table 8.

SciFact assigns NEI when no cited evidence exists, inadvertently creating a structural signal any expressive model will learn. Zero-shot GPT-4o attempts genuine reasoning on NEI instances. This shortcut inflates macro-F1 for all fine-tuned models and has gone unreported in prior work.

4.3 Bidirectional Out-of-Domain Generalization

Table 2 presents the full cross-dataset evaluation.

SciFact → HealthVer. All SciFact-trained models suffer catastrophic degradation as macro-F1 drops from 86-88% in-domain to 36-44% OOD. NEI collapse is most severe as HealthVer NEI has non-empty, topically relevant evidence and the absence shortcut does not fire. Table 2 shows the results. Qwen2.5-3B is the most NEI-resilient SciFact-trained model, suggesting broader pre-training provides some generalization advantage.

HealthVer → SciFact. Reverse direction tells a strikingly different story. Mistral-7B trained on just 1,008 HealthVer examples achieves 69.3% macro-F1 OOD on SciFact with 74.3% NEI F1, far above any SciFact-trained model on HealthVer. Remarkably, this outperforms BioLinkBERT trained on the full HealthVer set (60.8%), despite using 10× less

Model	Cost/1K	Fine-tune	vs. GPT-4o
GPT-4o (API)	$1.3000	—	1.0×
GPT-5 (API)	$0.5750	—	2.4× cheaper
Phi-3 (QLoRA)	$0.0292	$0.35	44.5× cheaper
Qwen2.5 (QLoRA)	$0.0292	$0.35	44.5× cheaper
Mistral (QLoRA)	$0.0292	$0.35	44.5× cheaper
BioLinkBERT	$0.0292	$0.17	44.5× cheaper

Table 3: Inference cost per 1,000 preds and one-time fine-tuning cost (T4 GPU). API pricing as of early 2026.

training data. BioLinkBERT at matched size shows meaningful transfer and achieves 53.3% macro-F1 and 67.1% NEI F1. Together, these results demonstrate that training on structurally sound data enables robust cross-domain transfer and model architecture matters more than data quantity.

Asymmetric OOD generalization. Symmetric degradation would indicate distributional shift; instead, collapse is strictly one-directional. The asymmetry holds at matched training sizes (1,008 examples), ruling out data quantity. Models trained on genuine epistemic reasoning (HealthVer) transfer; models trained on a structural proxy (SciFact) do not.

Per-class analysis. REFUTES F1 degrades most severely OOD in both directions: BioLinkBERT (HealthVer$_{full}$→SciFact) achieves only 29.1% despite strong SUPPORTS (77.9%) and NEI (75.3%) transfer. Directional contradiction detection requires domain-specific reasoning that generalizes poorly across biomedical claim types. This is a critical gap given that refutation detection is essential for any clinical task.

GPT-4o and GPT-5: paradoxical OOD behavior. GPT-4o achieves the highest HealthVer NEI F1 (63.1%) among zero-shot models despite the lowest SciFact NEI F1 (82.3%). Its in-domain weakness reflects genuine uncertainty reasoning that transfers. GPT-5 shows more pronounced NEI over-prediction: 95.5% NEI recall on SciFact but only 8.8% SUP recall on HealthVer, suggesting over-cautious RLHF training that predicts NEI regardless of evidence. In-domain scores are poor proxies.

4.4 Cost Analysis

GPT-4o costs $1.30 per 1,000 predictions versus $0.03 for local open models (44.5× drop). The $0.35 one-time fine-tuning cost amortizes to negligible overhead, making fine-tuned small models strictly preferable on both performance and cost grounds. For OOD tasks, training data choice is critical. SciFact-trained models collapse OOD while HealthVer-trained models transfer robustly, achieving strong cross-domain performance.

4.5 Qualitative Error Analysis

We examined 48 cases where Mistral-7B QLoRA is correct and GPT-4o is wrong, and 40 reverse cases, on the SciFact test set.

Fine-tuning wins. Mistral-7B outperforms GPT-4o on NEI: for "Statins increase blood cholesterol," GPT-4o predicts REFUTES from pharmacological knowledge while Mistral correctly withholds judgment given no retrieved evidence. Fine-tuning teaches models to attend to evidence rather than prior knowledge.

GPT-4o wins. GPT-4o retains an edge on directional REFUTES: for "The risk of male prisoners harming themselves is ten times that of female prisoners," the evidence states the opposite direction. GPT-4o correctly predicts REFUTES; Mistral latches onto "ten times" without detecting the subject inversion. Mistral nonetheless closes this gap vs. Phi-3-mini (11 GPT-4o wins vs. 29), suggesting larger capacity helps with contradiction detection.

Shared failure modes. Both models err on claims where evidence describes a related but experimentally distinct condition, and on logical negations introduced by experimental targeting. The annotation ambiguity is documented in SciFact inter-annotator studies (Wadden et al., 2020).

5 Conclusion

Fine-tuning small open-weight LLMs via QLoRA on just 1,008 examples surpasses both GPT-4o and GPT-5 on SciFact and HealthVer at a fractional cost, establishing parameter-efficient fine-tuning as a practical alternative to APIs. Through extensive OOD evaluation, we show both the promise and limits of this approach: HealthVer-trained models transfer robustly to SciFact, outperforming BioLinkBERT trained on 10× more data, while SciFact-trained models collapse cross-domain due to a structural artifact in SciFact. REFUTES remains the hardest class to transfer in both directions, pointing to cross-domain contradiction detection as an open challenge. Our findings motivate structural auditing of biomedical NLI benchmarks and bidirectional evaluation as standard practice.

Limitations

Our bidirectional experiments use matched 1,008-example training sets, but scaling to full HealthVer data would further improve in-domain performance, as evidenced by BioLinkBERT's 81.9% ceiling vs. 62.4% at matched size. We evaluated GPT-5 as an additional zero-shot baseline; its lower performance (77.9% SciFact, 42.4% HealthVer) confirms that QLoRA fine-tuning advantages hold against the latest proprietary systems. Cost estimates reflect T4/A100 and API pricing as of early 2026. We evaluate English claims only; whether analogous structural shortcuts exist in PubHealth, MedNLI, or other benchmarks remains an open question. Data augmentation and debiasing strategies for improving cross-domain robustness are left for future work.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Amin, Jyoti Aneja, Ahmed Awadalla, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Beber, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Miloš Košprdić, Adela Ljajić, Darija Medvecki, Bojana Bašaragin, and Nikola Milošević. 2024. Scientific claim verification with fine-tuned NLI models. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)*, pages 15–25.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhoujun Li, Dacheng Li, Eric Xing, and 1 others. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Label Normalization

Refer to Table 4.

Raw output variants	Label
SUPPORTS, SUPPORT, SUPPORTED, ENTAILMENT	SUPPORTS
REFUTES, REFUTE, REFUTED, CONTRADICT(S)	REFUTES
NEI, NOT_ENOUGH_INFO, NOT ENOUGH INFO, NEUTRAL	NEI
(unrecognized)	NEI (fallback)

Table 4: Label normalization mapping

B Prompt Templates

All instruction-tuned LLMs use the following prompt. Phi-3-mini uses ChatML format; Mistral-7B and Qwen2.5 use `[INST]/[/INST]` format.

```
System: You are a biomedical claim verification expert. Given a claim and an evidence passage from a scientific abstract, determine whether the evidence SUPPORTS the claim, REFUTES the claim, or whether there is NOT ENOUGH INFO. Respond with exactly one of: SUPPORTS, REFUTES, NEI.

User: Claim: {claim}
Evidence: {evidence}
What is the verdict?

Assistant: {label}
```

For GPT-4o, only the system and user turns are used at inference. For fine-tuning, the full exchange including the gold label is the SFT training target. For SciFact NEI examples, the evidence field is empty (`Evidence: `), the structural signal described in Section 4.2. We run GPT-4o with temperature 0 and GPT-5 with minimal reasoning.

C Dataset Statistics

Tables 5, 6 present dataset split sizes and label distributions. Tables 7, 8 provide representative input examples and evidence field statistics respectively

Dataset	Train	Val	Test	Total
SciFact	1,008	253	450	1,711
HealthVer (subset)	1,008	—	1,823	—
HealthVer (full)	10,590	1,917	1,823	14,330

Table 5: Dataset splits. SciFact uses stratified 80/20 train/val split; official dev set as test. HealthVer subset sampled with random_state=42 to match SciFact training size.

Dataset (train)	SUP%	REF%	NEI%
SciFact	48.8	27.1	24.1
HealthVer	35.7	22.8	41.5

Table 6: Label distributions. HealthVer's higher NEI proportion and structurally distinct NEI definition contribute to the generalization asymmetry.

Dataset	Claim	Evidence	Label
SciFact	Statins increase blood cholesterol.	*(empty)*	NEI
SciFact	Metformin reduces HbA1c in diabetic patients.	Metformin significantly reduced HbA1c levels compared to placebo...	SUP
SciFact	Aspirin prevents colorectal cancer.	Aspirin use was associated with reduced risk...	REF
HealthVer	Vitamin C prevents the common cold.	Some studies suggest marginal effects on duration but evidence remains inconclusive...	NEI
HealthVer	Exercise reduces depression symptoms.	Aerobic exercise significantly reduced depressive symptoms across trials...	SUP

Table 7: Representative input examples from SciFact and HealthVer. SciFact NEI examples have empty evidence fields by construction; HealthVer NEI examples contain present-but-inconclusive evidence, eliminating the absence shortcut. Evidence text truncated for brevity. This structural difference explains the asymmetric OOD generalization reported in Section 4.3.

Label	Count	Mean chars	Zero-length
NEI	243	0.0	243/243 (100%)
SUPPORTS	492	214.5	0/492 (0%)
REFUTES	273	227.4	0/273 (0%)

Table 8: Evidence field length by label in SciFact training split. All NEI examples have empty evidence; no SUP/REF example does. This holds in val and test splits as well.

D Hyperparameter Grid Search

We perform multi configuration grid search over $r \in \{8, 16, 32\}$ and lr $\in \{10^{-4}, 2\times10^{-4}, 5\times10^{-4}\}$ using Mistral-7B on SciFact validation set. Selected config (r=16, lr= 2×10^{-4}) in bold.

E Hyperparameter Configuration

We list down the hyperparameters used for fine-tuning the LLMs in Table 10.

r	LR	Val Macro-F1
16	$\mathbf{2\times10^{-4}}$	**88.4**
8	5×10^{-4}	87.4
16	5×10^{-4}	85.7
32	5×10^{-4}	84.6
8	2×10^{-4}	81.3
32	2×10^{-4}	79.0
16	1×10^{-4}	77.8
32	1×10^{-4}	75.7
8	1×10^{-4}	74.5

Table 9: Hyperparameter grid search results on SciFact validation set (Mistral-7B). Selected configuration in bold. Higher learning rates generally outperform lower ones; r=16 at lr= 2×10^{-4} achieves the best validation macro-F1 of 88.4%.

Hyperparameter	Value
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Quantization	NF4, double quantization
Learning rate	2×10^{-4}
LR schedule	Cosine, warmup ratio 0.05
Effective batch size	16 (batch 4, grad. accum. 4)
Max sequence length	1,024 tokens
Training epochs	3
Precision	bf16 (fp16 fallback on T4)
Optimizer	Paged AdamW 8-bit
Checkpoint selection	Best validation macro-F1
Trainable params (Phi-3)	8.9M / 3.83B (0.23%)
Trainable params (Mistral/Qwen)	41.9M / 7.29B (0.58%)
LoRA target modules:	
`q_proj, k_proj, v_proj, o_proj,`	
`gate_proj, up_proj, down_proj`	

Table 10: QLoRA configuration, identical across all architectures and datasets.

F Training Efficiency

We measure training efficiency as a function of training time and cost. All models were trained on Google Colab Pro. Phi-3-mini, Qwen2.5-3B, and BioLinkBERT were trained on a T4 GPU (16 GB); Mistral-7B required an A100 (40 GB) due to memory requirements at 7B scale. Times are reported for SciFact (1,008 examples, 3 epochs); HealthVer subset training is approximately equivalent. The detailed breakdown is in Table 11.

G McNemar's Test Details

McNemar's test statistic: $\chi^2 = (b_{01} - b_{10})^2/(b_{01} + b_{10})$, where b_{01} = model A correct, B wrong; b_{10} = model B correct, A wrong. Evaluated on SciFact test set (450 examples).

Model	GPU	Time	Cost
Phi-3-mini	T4 (16 GB)	∼50 min	$0.15
Qwen2.5-3B	T4 (16 GB)	∼70 min	$0.21
Mistral-7B	A100 (40 GB)	∼35 min	$0.35
BioLinkBERT	T4 (16 GB)	∼15 min	$0.04

Table 11: Training times and costs per model (SciFact, 1,008 examples, 3 epochs). T4: ≈$1.76 CU/hr; A100: ≈$10–15 CU/hr.

Model A	Model B	b_{01}	b_{10}	χ^2	p
Mistral QLoRA	GPT-4o	48	40	0.56	0.46
Mistral QLoRA	Phi-3 QLoRA	36	30	0.38	0.54

Table 12: McNemar's test on SciFact test set. Neither comparison is significant at $p < 0.05$.

H Qualitative Error Analysis: Extended Examples

Claim	Gold	Mistral	GPT-4o
Pattern A: Fine-tuned wins on NEI			
Statins increase blood cholesterol.	NEI	NEI	× REF
Venules have larger lumen than arterioles.	NEI	NEI	× SUP
Pattern B: GPT-4o wins on REFUTES			
Male prisoner self-harm risk is 10× female.	REF	× SUP	REF
Nicotine combo therapy yields significantly higher abstinence.	REF	× SUP	REF
Pattern C: Mistral wins on REFUTES (new vs. Phi-3)			
NIV use should be decreased for poor responders.	REF	REF	× NEI
Tirasemtiv has no effect on fast-twitch muscle.	REF	REF	× NEI
Pattern D: Both wrong			
TNFAIP3 is a tumor suppressor in glioblastoma.	REF	× SUP	× SUP
Low nucleosome occupancy correlates with low methylation.	REF	× SUP	× SUP

Table 13: Representative error examples. REF = REFUTES, SUP = SUPPORTS.

A. Fine-tuned models withhold judgment without evidence; GPT-4o applies world knowledge.

B. GPT-4o detects subject inversions and non-significant p-values; Mistral closes this gap vs. Phi-3-mini (11 wins vs. 29).

C. Mistral infers directional implications from indirect evidence where GPT-4o hedges to NEI.

D. Both models fail on logical negations and directional correlation signs (Wadden et al., 2020).

I Confusion Matrices

Figure 1 visualizes the NEI shortcut directly. BioLinkBERT trained on SciFact achieves perfect

NEI separation in-domain (left). The NEI column is completely clean with zero misclassifications. When evaluated on HealthVer OOD (right), the model predicts NEI exactly once across 727 true NEI examples, instead routing all NEI instances to SUPPORTS (647) or REFUTES (79). This collapse is the behavioral signature of shortcut learning: the absence signal that perfectly predicted NEI in SciFact is absent in HealthVer, and the model has learned nothing else.

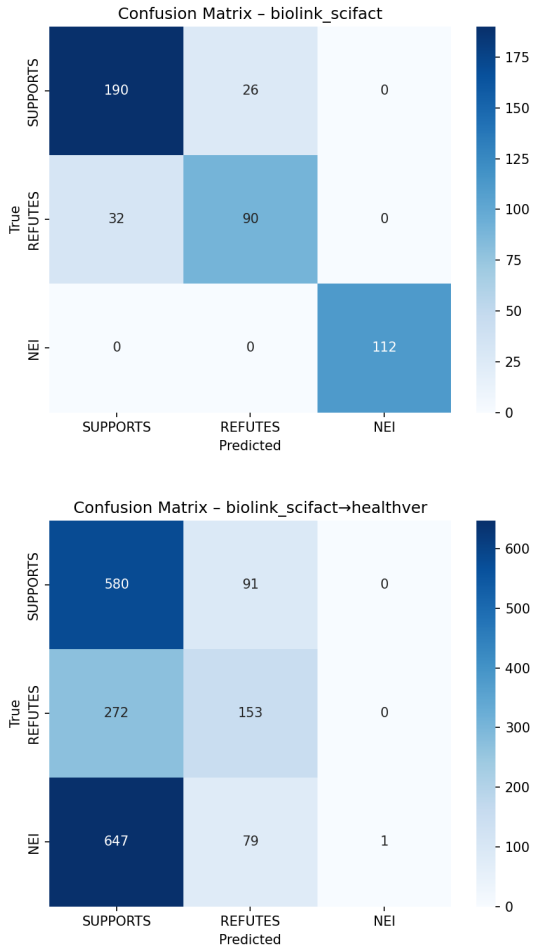


Figure 1: Confusion matrices for BioLinkBERT trained on SciFact, evaluated in-domain (top) and OOD on HealthVer (bottom). In-domain NEI is perfectly classified (112/112); OOD NEI is nearly never predicted (1/727), with true NEI examples overwhelmingly misclassified as SUPPORTS. This is the behavioral signature of structural shortcut learning.

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

Remy François
Parallia Healthcare
francois.remy@parallia.eu

Abstract

Reliable biomedical and clinical retrieval requires more than strong ranking performance: it requires a practical way to find systematic model failures and curate the training evidence needed to correct them. Late-interaction models such as ColBERT (Khattab and Zaharia, 2020) provide a first solution thanks to the interpretable token-level interaction scores they expose between document and query tokens. Yet this interpretability is shallow: it explains a particular document–query pairwise score, but does not reveal whether the model has learned a clinical concept in a stable, reusable, and context-sensitive way across diverse expressions. As a result, these scores provide limited support for diagnosing misunderstandings, identifying irreasonably distant biomedical concepts, or deciding what additional data or feedback is needed to address this. In this short position paper, we propose Diagnosable ColBERT, a framework that aligns ColBERT token embeddings to a reference latent space grounded in clinical knowledge and expert-provided conceptual similarity constraints. This alignment turns document encodings into inspectable evidence of what the model appears to understand, enabling more direct error diagnosis and more principled data curation without relying on large batteries of diagnostic queries.

1 Introduction

Trustworthy artificial intelligence in high-stakes domains increasingly requires transparent reporting of model capabilities, limitations, and failure modes (World Health Organization, 2021; European Parliament and Council of the European Union, 2017; Oberst et al., 2024). Finding systematic model failures is a central requirement for deploying retrieval systems in clinical and biomedical settings. When a retrieval model confuses nearby concepts, overgeneralizes across contexts, or fails to encode clinically salient negation, the consequence is not merely a lower ranking metric, but a weaker basis for downstream decision making and a poorer signal for curating corrective training data (Harkema et al., 2009; Lee et al., 2010; Wang et al., 2020). For this reason, debugging should ideally reveal not only that a model failed on a particular document–query pair, but also *what* the model appears to understand, *where* that understanding breaks down, and *which* additional examples are needed to improve it.

Late-interaction models in the style of ColBERT are appealing in this regard because they expose fine-grained interaction scores between query and document tokens (Khattab and Zaharia, 2020). However, this evidence remains retrospective and local: it helps explain why a given query matched a given document, but not whether the model has learned clinically meaningful distinctions robustly across paraphrases, contexts, and compositional variants (Huang and Baldwin, 2023; Kang et al., 2025). Explaining a match is not the same as gauging and diagnosing misunderstandings.

In clinical retrieval, the most consequential failures often depend on context-sensitive factors such as negation, temporality, uncertainty, experiencer, historical status, and the distinction between mentioning a test and asserting a condition. These are precisely the kinds of distinctions that may be weakly expressed in local match scores while remaining decisive for retrieval quality.

We therefore argue that checking model understanding requires interpretation at multiple levels of contextualization. Many biomedical meanings are only partially specified at the token level and become clearer as progressively more context is introduced (Lee et al., 2010). For example, a model that appears reasonable on *cat* may still fail on *cat scratch disease*. Moreover, paragraph-level context can further influence whether a mention is current, historical, speculative, negated, or even concerns another person (Harkema et al., 2009; Wang et al., 2020). Interpretability must therefore extend

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 713–717
July 3-4, 2026 ©2026 Association for Computational Linguistics

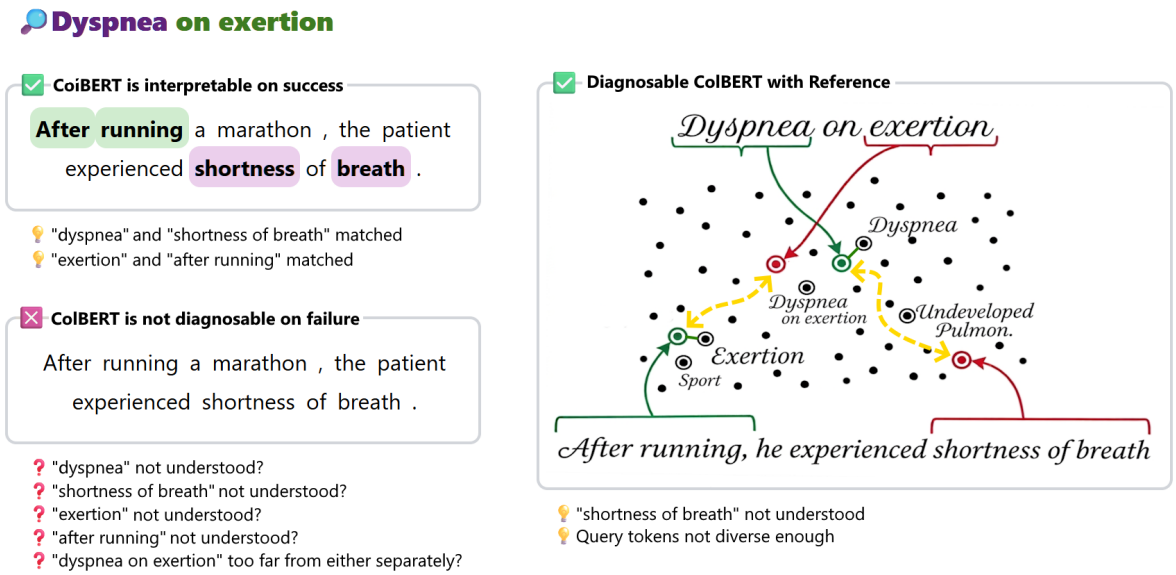


Figure 1: Standard ColBERT-style late interaction produces interpretable query–document scores, but provides little guidance when no meaningful match is found for a token, as illustrated on the left. By contrast, augmenting the retriever with a reference latent space, as on the right, allows testers of Diagnosable ColBERT models to diagnose failures more quickly and more actionably by separating latent-space semantics from in-context understanding.

beyond isolated token matches toward a graded view of understanding across term, sentence, and paragraph-level alignment. Accordingly, we treat diagnostic alignment as multi-factorial: it should reveal term-level concept identity, phrase-level composition, context-level qualifiers, and broader discourse elements that alter clinical interpretation.

This broader view of interpretation also motivates the use of a structured latent space rather than a fixed inventory of labels. If the embedding space is shaped to respect concept-level and sentence-level similarity constraints, then it can represent not only previously named clinical concepts, but also contextually enriched and partially novel combinations of them (Campbell et al., 2014). Prior work also suggests that such spaces are feasible to construct in practice, whether by grounding biomedical representations in ontological definitions and relations as in BioLORD (Remy et al., 2024), by aligning dense representations to structured label spaces (Decorte et al., 2025), or by grounding annotated mentions in a dynamic model-generated concept space when the concept inventory is large and evolving (Stepanov et al., 2026).

In this publication, we therefore propose Diagnosable ColBERT, a framework that leverages such a reference latent space learned from expert-specified similarity constraints over clinical concepts, and aligns late-interaction token embeddings to it via pre-projection adapters. The central proposal of this paper is therefore not a new retrieval scoring rule, but a new diagnostic lens: late-interaction representations should be interpreted against a clinically grounded reference space rather than only through pairwise match scores. The goal is not merely to explain a retrieval score after the fact, but to make document encodings themselves interpretable as evidence about what the model appears to know. This creates a path toward diagnosing misunderstandings directly from encoded documents, identifying confusable concepts, and curating training data in a more principled way. More generally, we advocate a shift from retrospective score inspection to concept-grounded diagnosis of model understanding. Such representations could support both direct human inspection and more automated auditing workflows.

2 Diagnostic Framework

Diagnosable ColBERT is organized around a pre-existing reference latent space that serves as a domain-specific diagnostic scaffold. This latent space needs to accommodate both concept names, clinical sentences, and paragraphs; see for example BioLORD (Remy et al., 2024). The purpose of this space is to make contextual token representations clinically legible: not only in terms of term-level concept identity, but also in terms of local composition and context-level qualifiers such as negation, temporality, uncertainty, or experiencer. Instead of only asking which query token matched which document token, we ask what clinically relevant factors a contextualized representation appears to encode and how they arrange geometrically.

In this view, alignment means mapping late-interaction token representations into a space where these factors can be inspected more directly. The retrieval representation should remain tied to this diagnosed representation, but need not be identical to it. A natural design is for retrieval embeddings to be learned as a lower-dimensional downprojection of the diagnosed representation, so that retrieval can reweight clinically relevant factors for ranking efficiency without discarding the richer structure needed for diagnosis. This keeps the framework centered on a simple idea: retrieval and diagnosis should stay coupled, but diagnosis should not be reduced to whatever geometry happens to be most convenient for fast search.

The specific architecture used to realize this idea is secondary to the paper's main claim. One can implement the diagnostic view with lightweight contextual adapters and task-specific projection heads; our essential contribution is the diagnostic framing itself: a clinically grounded reference geometry for inspecting what a late-interaction retriever appears to perceive from a context.

3 Practical Examples

A central limitation of standard ColBERT interpretability is that it is fundamentally relational: it explains why a query matched a document, but offers limited leverage when the goal is to determine why a relevant match failed to occur (see Figure 1). Diagnosable ColBERT is intended to address precisely this gap. Rather than treating retrieval failure as a single undifferentiated event, it seeks to identify which part of the representation pipeline failed to preserve the clinically relevant meaning.

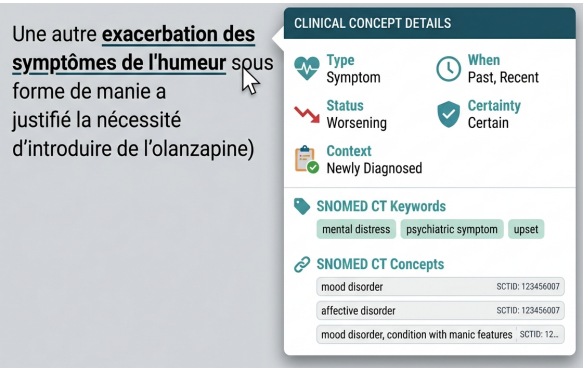


Figure 2: Proposed debugging interface for Diagnosable ColBERT. The interface exposes token-level query and document representations together with their placement in the reference latent space and the clinically meaningful concepts or qualifiers associated with nearby regions. This gives testers a practical way to distinguish weak interaction scores from deeper failures of concept grounding, abbreviation handling, or contextual interpretation, and to turn a retrieval miss into an actionable diagnosis.

Consider a case report retrieval system in which a tester issues the query *bartonellosis*. A relevant report is missed because the report mentions only *CSD*, the abbreviation for *cat scratch disease*. Standard query–document inspection establishes that the interaction between *bartonellosis* and *CSD* is too weak. But this finding alone is diagnostically incomplete. It does not tell us whether the query representation failed to capture the target disease family, whether the document representation failed to interpret the abbreviation correctly, or whether both encodings are deficient.

Diagnosable ColBERT resolves this ambiguity by grounding both sides in a reference latent space. The tester can inspect whether *bartonellosis* is already positioned near the relevant disease concept and, separately, whether *CSD* is mapped into that same region. If *bartonellosis* is correctly grounded but *CSD* is not, then the root cause is not weak interaction per se, but document-side abbreviation understanding. If both are misplaced, the problem lies deeper, at the level of concept representation. This kind of distinction is difficult to obtain from pairwise interaction scores alone, yet it directly guides remediation.

When the problem lies in abbreviation grounding, the natural intervention is to curate examples that explicitly connect abbreviated, expanded, and synonymous forms. When the problem lies in broader concept placement, the remedy instead concerns the shaping of the underlying semantic space.

The same diagnostic logic extends beyond query-conditioned retrieval. Some failures can be identified directly from the document encoding itself. Consider the utterance: *Are you allergic to any of the following drugs? — Yes, to ranitidine*. Here the relevant question is whether the token representation of *ranitidine* reflects not only the medication identity, but also the allergic relation introduced by context. If the encoding remains close to the drug concept while failing to capture the neighboring context of allergy, adverse reaction, or intolerance, then the model has encoded the entity but not the clinically decisive context.

This is a query-free diagnostic signal: the tester can identify the failure before retrieval is even attempted. Again, the diagnosis suggests a targeted intervention, namely the addition of training examples in which drug mentions appear indirectly in contexts related to allergy or adverse-reactions, and ensure that this context is getting properly picked up by the encoder model.

4 Real-World Usage

In our experience, debugging user interfaces for Diagnosable ColBERT models[1] have already proven useful as a practical way to inspect token-level phenomena and surface failures that would have been difficult to localize from ranking metrics alone. What made the interface valuable was that it helped spot remaining representational weaknesses that actually mattered in practice, including gaps in conceptual grounding, abbreviation and brand name handling, and preservation of clinical context.

Subjectively, this changed the development process. Instead of treating misretrievals as abstract evaluation outcomes, we were often able to trace issues back to recurring representational problems and then respond with more targeted curation and model changes. These observations informed the progression from ClinicalEncoder25 to ClinicalEncoder26AM and beyond, even if this workflow remains work in progress and should not be mistaken for a complete evaluation methodology.

5 Limitations and Scope

Diagnosable ColBERT is best understood as a diagnostic scaffold, not as a complete semantics of biomedical meaning. Its purpose is to make clinically relevant aspects of the model's internal organization more legible to a tester, not to claim that every retrieval-relevant distinction can be exhaustively captured in a single reference geometry. Any such space reflects design choices about which concepts, relations, and contextual qualifiers are made salient, and its usefulness depends on being well shaped by domain knowledge, expert constraints, and clinically meaningful similarity structure. The proposal therefore improves diagnosability without by itself guaranteeing retrieval quality: diagnostic alignment complements retrieval evaluation by making failure modes more interpretable and actionable, but useful systems will still need strong ranking behavior as well.

6 Conclusion

In this paper, we have argued that debugging late-interaction retrievers requires more than inspecting token-level query–document scores after a failure has already occurred. Our central claim is that retrieval models become substantially more diagnosable when their token representations are aligned to a clinically grounded reference latent space that makes concept grounding, contextual qualifiers, and representational gaps directly inspectable.

Diagnosable ColBERT is a concrete step in this direction. Rather than treating retrieval as a black box whose failures must be inferred indirectly from weak matches and downstream metrics, it provides a practical framework for localizing why a miss occurred and for inspecting difficult document segments even before a query is issued.

Much remains to be refined, especially in constructing strong reference spaces and turning diagnoses into robust improvement workflows. Even so, we view the direction as already promising in practice: clinically grounded diagnostic representations complement ranking evaluation by making retrieval behavior more legible and actionable.

[1] Public demos: http://demo26.parallia.eu/ and http://text2json.parallia.eu.

References

Walter S. Campbell, James R. Campbell, William W. West, James C. McClay, and Steven H. Hinrichs. 2014. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. *Journal of the American Medical Informatics Association*, 21(5):885–892.

Jens-Joris Decorte, Jeroen Van Hautte, Chris Develder, and Thomas Demeester. 2025. Efficient text encoders for labor market analysis. *IEEE Access*, 13:133596–133608.

European Parliament and Council of the European Union. 2017. Regulation (EU) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union, L 117, pp. 1–175. Adopted April 5, 2017; published May 5, 2017.

Hendrik Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Yiran Huang and Timothy Baldwin. 2023. Robustness tests for automatic machine translation metrics with adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4649–4675.

Junmo Kang, Yunhyeok Ro, Junsie Heo, and Minjoon Seo. 2025. TRIAL: Token relations and importance aware late-interaction for accurate text retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15404–15427.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Dennis H. Lee, Francis Y. Lau, and Hue Quan. 2010. A method for encoding clinical datasets with SNOMED CT. *BMC Medical Informatics and Decision Making*, 10:53.

Michael Oberst, Davis Liang, and Zachary C. Lipton. 2024. Pioneering the science of AI evaluation. Whitepaper, published September 19, 2024; updated August 7, 2025.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. 2026. The million-label NER: Breaking scale barriers with GLiNER bi-encoder. *Preprint*, arXiv:2602.18487. Introduces GLiNKER, a modular framework for large-scale entity linking.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: A resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

World Health Organization. 2021. Ethics and governance of artificial intelligence for health: WHO guidance. Published June 28, 2021.

Developing Literature Annotation Guidelines for Representing Normal Physiology in Biolink-Compatible Knowledge Graphs

Madeline Bittner, Willie Rogers, Dina Demner-Fushman,
Richard H. Scheuermann, Matthew Diller
Division of Intramural Research
National Library of Medicine, Bethesda, MD
firstname.lastname@nih.gov

Abstract

Much of our knowledge about anatomy and physiology is found in text format in research papers and medical textbooks. For an information system to have access to this knowledge, extracting and translating it into a computable format that can be stored in an ontology or knowledge graph is advantageous. Unfortunately, existing text mining corpora, which are needed to train and evaluate data mining models, are old and consist almost entirely of research papers, which rarely contain complete information needed to capture complex normal physiological processes and, subsequently, understand the pathophysiology of a disease. As a first step to filling in this gap, we have developed a guide for annotating medical textbooks for physiological events and entities involved in these events. In addition to providing our guidelines and describing the guideline development process, we analyze the coverage of normal physiology in existing ontologies.

1 Introduction

Anatomy and physiology are central components of biological systems, with knowledge of the relationship between the structures and functions of organisms and their parts driving most of what we know about health and disease. Most knowledge about healthy anatomy and physiological functions exists as text in medical textbooks and research papers, which is not easily accessible in a structured form for querying, analysis, and visualization. This lack of structured, computable knowledge representation has hindered the translation of large-scale multi-omics datasets into new knowledge about disease mechanisms and potential treatments. In addition, many ontologies and knowledge graphs (KGs) contain large gaps in coverage as they tend to lag behind novel research findings in terms of being up-to-date, and rely largely on manual curation (Li et al., 2023). To scale up curation efforts, ontology and KG developers will need to incorporate semi-automated tools and workflows for extracting knowledge about anatomical entities and physiological events from biomedical text.

Here we describe an annotation guide that can be used as a reference for annotating anatomical entities and physiological events in medical textbooks. Included in these guidelines are the scope of the annotation, a description of each entity and event of interest, and the reasons for why certain annotation decisions were made. To test the feasibility of translating textual physiology information into a structured format, we assess the coverage of extracted entities and events in existing biomedical ontologies.

2 Background

Currently, several tools exist to create structured representations of knowledge in ontologies and KGs that could be applied to anatomy and physiology, including Biolink (Unni et al., 2022), which aims to provide an upper-level model for interoperable KGs. However, there is an absence of up-to-date text mining corpora to support the extraction of anatomy and physiology information from text. The most recent annotated corpora—the Anatomical Entity Mention (AnEM) corpus (Ohta et al., 2012), the extended Anatomical Entity Mention (AnatEM) corpus (Pyysalo and Ananiadou, 2013), and the Multi-Level Event Extraction (MLEE) corpus (Pyysalo et al., 2012)—that broadly focus on mentions of anatomical entities and physiological events in the context of both healthy and diseased organisms, largely consist of annotated abstracts and full-text research publications. Although these corpora are annotated with ontology term identifiers, the length of time since these annotations were performed would warrant reannotation to ensure that active and the most precise terms are used. Additionally, since the annotated documents in these corpora consist mostly of research papers,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 718–728
July 3-4, 2026 ©2026 Association for Computational Linguistics

their structure and content will likely differ from the semantic and syntactic structures found in a medical textbook. Beyond the limitations of existing corpora, existing guidelines used for biomedical information extraction, including i2b2 (Uzuner et al., 2011) for clinical de-identification, GENIA (Kim et al., 2003) for biochemical entities in molecular biology literature, and CINEX (Reichenpfader et al., 2025) for high-level metadata extraction from clinical trial text, were not designed to capture normal physiology entities, as they are either too metadata-focused or operate at a biochemical level that does not translate to physiology annotation. To our knowledge, no annotation guidelines currently exist that are specific to normal human physiology, representing a gap that this work addresses.

3 Methods

In this exploration of approaches to annotation of physiological processes and events at several anatomy levels, we focus on annotating events that involve glucagon-like peptide 1 (GLP-1), which plays an essential role in processing nutrients after food ingestion. The effects of GLP-1 are also mimicked by agonistic drugs, such as exenatide (Byetta), liraglutide (Victoza), dulaglutide (Trulicity) and semaglutide (Ozempic/Wegovy), used for blood glucose control and weight management.

3.1 Annotation Schema Development

The annotation guide was developed using passages from the *Guyton and Hall Textbook of Medical Physiology* (14th Edition) (Hall et al., 2021), well-established reference in biomedical education and clinical medicine. This textbook was selected because it contains medically complex language representative of normal physiology text, is well-structured, and reflects the type of text that would be encountered when annotating other normal physiology domains beyond the GLP-1 use case. Passages were drawn specifically from sections covering GLP-1 normal physiology, with 17 paragraphs selected for annotation.

Three annotators with expertise in biomedical informatics annotated each text independently. Annotations were conducted iteratively, with label definitions and disambiguation rules continuously refined to improve clarity and ensure generalizability beyond GLP-1 physiology to other normal human physiology domains. Comprehensive coverage of relevant biological concepts while remaining broad enough to generalize to other normal human physiology processes. Text was considered out of scope if it describes disease pathways, pertains to processes at a purely biochemical level, or does not relate to normal adult human physiology. Entity, event, and attribute labels and their definitions are provided in appendix Table A1, and were informed by prior work on manual annotation of anatomical and physiological features in medical text (Ohta et al., 2012; Pyysalo and Ananiadou, 2013; Pyysalo et al., 2012). Counts for the number of occurrences of each entity, event, and attribute labels across the 63 annotated sentences from the 17 passages are provided Table 1. Target-Arg relations are used to formally link the events to their associated entities or other events, where the direction of the relation is defined as the source term to the target term. To reduce cross-sentence complexity, each sentence was annotated as independently as possible, with Target-Arg relations restricted within sentence boundaries when possible; the coreference attribute was used to capture topics that carries across sentences. Relationship serves as the catch-all event label, applied to any word or phrase that describes how entities are connected, interact, or influence one another. Three annotators independently annotated the corpus using BRAT (Stenetorp et al., 2012), as shown in Figure 1, with disagreements resolved through discussion and a predefined list of common multi-word expressions used to reduce complexity and increase annotation efficiency (appendix Table A2).

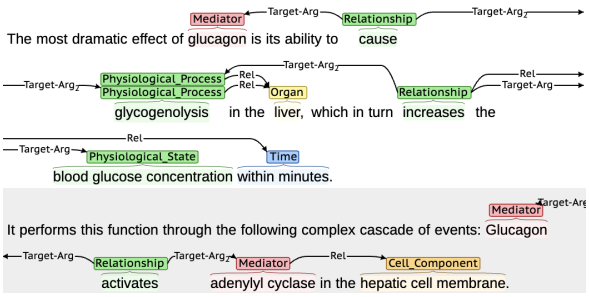


Figure 1: Example BRAT annotation.

3.2 Named Entity Recognition

To evaluate the feasibility of automated named entity recognition (NER) and to inform ontology selection for future work, we assessed the coverage of our annotation guide across a broad range of biomedical ontologies using the EMBL-EBI Ontology Lookup Service (OLS) (Côté et al., 2006) and

Category	Label	Count (N=420)
Entity	Mediator	33 (7.8)
	Measurement	24 (5.7)
	Chemical	22 (5.2)
	Organ	15 (3.5)
	Substrate	12 (2.86)
	Time	11 (2.6)
	Tissue	9 (2.1)
	Cell	7 (1.6)
	Organ System	2 (0.4)
	Cell Component	1 (0.2)
Event	Relationship	74 (17.6)
	Physiological State	50 (11.9)
	Physiological Process	47 (11.1)
Attribute	Rel	98 (23.3)
	Coreference	8 (1.9)
	is a	7 (1.6)

Table 1: Total counts (%) for each entity, event, and attribute label with percentages rounded to one decimal place.

the Unified Medical Language System (UMLS) (Bodenreider, 2004). A distinct list of Entities (N=29), Physiological Processes (N=23), Physiological States (N=13), and Relationships (N=55) was derived from our manual annotations. Multiword expressions and fragments were excluded prior to ontology lookup, as their structure is not directly represented in standard biomedical ontologies, though their constituent terms are covered individually. The terms were queried via the EMBL-EBI API (European Bioinformatics Institute (EMBL-EBI), 2026) and matched against available ontologies using exact string lookup. Coverage was calculated as the proportion of distinct manually annotated terms present in each ontology. The Entity labels *Time* and *Measurement* were excluded from this analysis, as these labels do not correspond to concepts represented within standard biomedical ontologies.

3.3 Relation Extraction

To evaluate the feasibility of automated relation extraction, we assessed whether Stanza (Qi et al., 2020) dependency and constituency parsers can structurally encode relationship types defined in our annotation guide. Both parsers were run on the 17 text paragraphs, with multi-word expressions adjoined with underscores to prevent term splitting. Manual relation annotations were mapped to parser outputs according to each parser type's structural properties: dependency parser spans were reduced to syntactic head words; while constituency parser, spans were aligned to leaf node sequences with Part-of-speech (POS) tags normalized to phrase-level equivalents. Each relation was parsed independently at the sentence level to avoid full-document parse context effects. Across the 162 annotated relation pairs, 29 distinct dependency patterns and 6 distinct constituency patterns were identified as the reference set. Coverage was assessed at two levels. Corpus-wide coverage determined whether each distinct relation pattern appeared anywhere in the full parser output across all 17 paragraphs, assessed for the dependency parser via direct UPOS $\rightarrow$ deprel $\rightarrow$ UPOS arc matching, and for the constituency parser via phrase label pair matching *(e.g., VP Target-Arg NP)* within the same sentence tree. Individual relation coverage determined whether each manually annotated pattern was present in its source sentence. For the dependency parser, using direct arc lookup between entity head words in both directions; for constituency parser, using the lowest common ancestor method to identify the smallest constituent dominating both entity head words and comparing its label to the annotated phrase type. Of the 162 annotated relation pairs, 2 were removed from the dependency pattern set and 18 were removed from the constituency pattern set prior to individual relation coverage, as these instances could not be reliably aligned to their corresponding sentence-level parse structures. Additionally, hierarchical agglomerative clustering (HAC) was performed to visually model and compare the relation pattern coverage of the dependency and constituency parsers, methods can be seen in appendix Figures A1 and A2.

4 Results

4.1 Inter Annotator Agreement

Mean pairwise F1 scores were computed to measure inter-annotator agreement, as shown in Table 2. Agreement for Entities, Relationships, which includes the labels Relationship, Physiological State, and Physiological Process, was measured at the character span level. For annotations with character span agreement, label agreement was then assessed. Agreement for Target-Arg relations was evaluated using the character spans for both the source and target terms, with label agreement then assessed for pairs that reached character span agreement. For Target-Arg label agreement, both the source and target labels were required to match. A token alignment leniency of two tokens was applied across all agreement levels. Attribute labels were excluded from inter-annotator agreement cal-

Section	Span	Label
Entities	0.626 (N=612)	0.814 (N=280)
Relationships	0.621 (N=809)	0.868 (N=365)
Target-Arg	0.354 (N=1202)	0.731 (N=260)

Table 2: Mean pairwise F1 agreement scores for span-level and label-level agreement across three annotators. N indicates matched spans/relations for label agreement.

culations, as their complexity and dependency on broader context make standard pairwise span-level evaluation unreliable. While span-level agreement is comparatively lower, this is expected given the nature of span boundary identification. Notably, label agreement among spans that all annotators agreed on is high, suggesting that annotators can reliably assign the correct label when a span is identified.

4.2 Ontology Coverage

Table A3 summarizes the coverage of each ontology from the OLS and UMLS. Among the OLS ontologies, OMIT ontology achieved the highest coverage for Entities (0.52), while NCIT achieved the highest coverage for Physiological Process, Physiological State, and Relationship (0.17, 0.38, and 0.24 respectively). Notably, PRIDE and INO shared the highest coverage of Physiological Process terms, and PRIDE also shared the highest coverage for Physiological State terms. Across all OLS ontologies, Physiological Process terms had the lowest overall coverage (0.17). In the UMLS, Relationship terms had the lowest coverage (0.11), while Entities had the highest (0.55). Overall, the UMLS achieved the highest coverage across all ontologies at 0.31, with NCIT second at 0.29.

4.3 Relation Coverage

Corpus-wide coverage was computed as the proportion of distinct manual annotation patterns found within the full parser output. The dependency parser achieved a corpus-wide coverage of 0.48, recovering 14 of 29 distinct manual relations, with a precision of 0.10, indicating that the 14 matched patterns account for only a small fraction of the 135 distinct patterns produced by the parser across the full corpus. The constituency parser achieved perfect corpus-wide coverage of 1.00, recovering all 6 distinct manual relations, with a precision of 0.19, meaning the 6 matched patterns represented a similarly small proportion of the 31 distinct patterns generated. At the individual relation level, the dependency parser covered 80 of the 160 eligible manual relation instances (0.50), and the constituency parser covered 89 of the 144 eligible manual relation instances (0.62). Results for the HAC dendrograms can be found in appendix Figures A1 and A2.

5 Discussion

Through this work, physiology annotation guidelines were developed to support corpus development for automatic named entity and relation extraction from medical text using NLP methods, towards the goal of knowledge graph triplet generation. The evaluation results reveal both promise and limitations of using these guidelines as a foundation for that task. While the guidelines provide a structure for both entity and relation annotation that can be generalized to cover a broad range of normal physiology, the feasibility of automated triplet extraction depends on the performance of both NER and relation extraction, each of which presents with different challenges.

In terms of NER, the ontology coverage evaluation demonstrated that no single ontology is sufficient to capture the full range of terms present across all categories. This is consistent with prior work showing that ontologies often provide only partial representations of a domain (Abad-Navarro et al., 2025). The UMLS exhibited the best overall coverage, however relationship terms showed lower coverage, likely reflecting that UMLS is primarily organized around biomedical entities rather than relational predicates. Additionally, these ontologies are generally not designed to capture higher-level, abstract representations, but instead focus on domain specific-concepts. As a result, although our terms are similar to those found in existing ontologies, they are often not represented because they are broader in scope and intended to capture relationships across a wide range of physiological contexts. For example, in the UMLS, the generalized relationship term *Prevents* is not explicitly represented as a standalone concept; however, more specific phrases such as *Chair prevents rising*, *Prevents risky behaviors*, and *Prevents skin barrier interruption* are included. These findings support prior work demonstrating that UMLS does not fully cap-

ture all concepts and relationships within a given domain (Yu et al., 1999). Therefore, our future work will explore combining multiple ontologies, as integrating resources has been shown to support more complete domain representation (Osman et al., 2021), alongside mapping strategies to improve entity coverage across all categories.

For relation extraction, the dependency and constituency parsers show different results in relation coverage. The dependency parser's slightly higher individual manual relation coverage (0.50) relative to its corpus-wide distinct relation coverage (0.48) suggests it performs better on more frequently occurring relation types found in our corpus. The constituency parser, despite achieving perfect corpus-wide distinct relation coverage, showed a notable drop at the individual relation level (0.61), driven largely by a high false positive rate. Nevertheless, the constituency parser shows greater promise as a basis for automated relation extraction, and prior work has demonstrated that constituency-based approaches can be effective when combined with rule-based filtering to reduce false positives (Jiang and Diesner, 2019). In future work, we aim to expand the text corpus to additional open access medical textbooks and explore how post-processing of constituency parser outputs can be used to extract relevant relations.

6 Conclusion

We present an annotation guide for anatomical entities and physiological events in biomedical textbooks, designed to support the automated construction of knowledge graph triples from free text. To access the feasibility of translating free medical text into a structured format, we annotated a selected text corpus and evaluated ontology coverage of our annotations across OLS ontologies and the UMLS. We evaluated the coverage of relation extraction by assessing the coverage of annotated relation pairs within parser outputs. Our results demonstrate that while existing ontologies and parsers provide partial coverage, significant gaps remain. These findings motivate future work in multi-ontology integration and post-processing strategies for parser outputs. This work represents a step toward making rich knowledge present in medical textbooks accessible in a computable format that can support development of clinical knowledge discovery platforms.

7 Limitations

This study is a pilot study utilizing a smaller sample size with text specific to our use case of GLP-1 physiology. This use case was selected because it was believed that a schema developed from it could be generalized and applied to other physiological functions, though this remains to be validated.

8 Ethics

This work includes limited use of content from *Guyton and Hall Textbook of Medical Physiology* (14th Edition) (Hall et al., 2021) for purposes of non-commercial research and computational analysis. The authors have conducted a good-faith assessment and believe this use is consistent with fair use principles (Association of College and Research Libraries, 2025).

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Francisco Abad-Navarro, Catalina Martínez-Costa, and Jesualdo Tomás Fernández-Breis. 2025. Ontology coverage analysis through language models. *Engineering Applications of Artificial Intelligence*, 162:112671.

Association of College and Research Libraries. 2025. Fair use. https://acrl.libguides.com/scholcomm/toolkit/fairuse. Accessed: April 17, 2026.

Daniel C Berrios, Sara C Jorgensen, Rachel K Shoop, Carlos De Los Santos, Danielle K Lopez, Alan E Wood, Gregory D Eley, Kathryn C Richards, Olga V Kieschnick, Sigrid S Reinsch, and 1 others. 2024. Increasing data discovery and re-use: The space life sciences ontology. In *Annual Meeting of the American Society for Gravitational and Space Research (ASGSR)*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Richard G. Côté, Philip Jones, Rolf Apweiler, and Henning Hermjakob. 2006. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7:97.

Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350.

European Bioinformatics Institute (EMBL-EBI). 2026. Ontology Lookup Service (OLS) API. Accessed: 2026-04-07.

John E. Hall, Michael E. Hall, and Arthur C. Guyton. 2021. *Guyton and Hall Textbook of Medical Physiology*, 14 edition. Elsevier.

Jingshan Huang, Christopher Townsend, Dejing Dou, Haishan Liu, and Ming Tan. 2011. Omit: a domain-specific knowledge base for microrna target prediction. *Pharmaceutical research*, 28(12):3101–3104.

Ming Jiang and Jana Diesner. 2019. A constituency parsing tree based method for relation extraction from abstracts of scholarly publications. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 186–191, Hong Kong. Association for Computational Linguistics.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Harry Li, Gabriel Appleby, Camelia Daniela Brumar, Remco Chang, and Ashley Suh. 2023. Knowledge graphs in practice: characterizing their users, challenges, and visualization opportunities. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):584–594.

James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118.

Todd Millecam, Austin J Jarrett, Naomi Young, Dana E Vanderwall, and Dennis Della Corte. 2021. Coming of age of allotrope: proceedings from the fall 2020 allotrope connect. *Drug Discovery Today*, 26(8):1922–1928.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36, Jeju Island, Korea. Association for Computational Linguistics.

Ibrahim Osman, Stefano F. Pileggi, Sadok Ben Yahia, and Gayo Diallo. 2021. An alignment-based implementation of a holistic ontology integration method. *MethodsX*, 8:101460.

Arzucan Özgür, Junguk Hur, and Yongqun He. 2016. The interaction network ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature. *BioData mining*, 9(1):41.

Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, and 1 others. 2022. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, 50(D1):D543–D552.

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2025. Improving the reporting quality of studies on information extraction from clinical texts: Protocol for the development of a consensus-based reporting guideline. *JMIR Research Protocols*, 14:e76776.

Winifred Sewell. 1964. Medical subject headings in medlars. *Bulletin of the Medical Library Association*, 52(1):164.

Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. 2007. Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, 40(1):30–43.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Deepak R Unni, Sierra AT Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J Harry Caufield, Paul A Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, and 1 others. 2022. Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and translational science*, 15(8):1848–1855.

U.S. National Library of Medicine. 2026. SNOMED CT International Edition, April 2026 Release. Release date: 2026-04-01; Accessed: 2026-04-07.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2011. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 18(5):552–563.

Danielle Welter, Simon Jupp, and David Osumi-Sutherland. 2018. Human cell atlas ontology. In *ICBO*.

Hong Yu, Carol Friedman, Andrey Rzhetsky, and Pauline Kra. 1999. Representing genomic knowledge in the umls semantic network. In *Proceedings of the AMIA Symposium*, pages 181–185. American Medical Informatics Association.

A Appendix

Category	Label	Definition	Examples
Entity	Cell	A material entity that is the basic structural unit of tissue.	Liver cells, alpha cells
	Cell Component	A subcellular structure or molecular complex that is associated with a cell and contributes to its function.	Nucleus, mitochondria
	Chemical	A molecule, substance, or compound involved in a physiological function where its specific role is unclear, unknown, or not relevant.	Blood glucose, glycogen in the liver
	Measurement	A data item that quantifies something.	140 mg, fivefold, one-third
	Mediator	A chemical that actively influences, regulates, or triggers a physiological reaction or process.	Glucagon
	Organ	A distinct anatomical structure made up of multiple tissue types that performs one or more biological functions.	Liver, heart
	Organ System	A group of distinct anatomical structures that work together to perform one or more biological functions.	Respiratory system
	Substrate	A chemical that is acted upon, consumed, or transformed in a biological process or reaction.	Blood glucose
	Time	A temporal descriptor for how long or what pattern a physiological process occurs.	Few minutes, within minutes, first hour
	Tissue	A group of similar cells organized together for a specific physiological function.	Blood, connective tissue
Event	Physiological Function	A biological activity or outcome carried out by cells, tissues, organs, etc. that contributes to normal body function.	Secretion, glucose uptake
	Physiological State	A condition, status, or internal environment of the body at a given time.	Blood glucose concentration, fasting, exhaustive exercise
	Relationship	A word or phrase that describes how two or more annotated entities are connected, interact, or influence each other within the physiological context.	Activates, decreases, rises, increases
Attribute	Coreference	When two entities refer to the same named entity.	Blood glucose (increase) *coreference* this (increase)
	is_a	One entity is a type or category of another.	GLP-1 *is_a* chemical
	Rel	When two entities are in relation to another.	Release *Rel* insulin

Table A1: Label definitions for the entities, events, and attributes included in the annotation schema. Contents within parenthesis are additional context not included in the named entity. Entities, Chemical, Substrate, and Mediator, were developed to distinguish molecules and compounds based on their functional role within the physiological context rather than chemical identity. Chemical serves as the catch-all label, substrate and mediator are more specific subcategories of Chemical. Three attribute labels, Coreference, is_a, and Rel, were developed to capture additional contextual information that cannot be fully expressed through entity or event labels alone.

Common Multi-Word Expressions
Blood glucose (concentration/level)
(Increase/Decrease) in blood glucose
(High/Low) concentrations
Effects of (Chemical/Mediator/Substrate)
(Chemical/Mediator/Substrate) secretion
(Chemical/Mediator/Substrate) function
(Chemical/Mediator/Substrate) concentration
Blood glucose buffer system
Enzyme system
Glucose (stimulus/concentration)
Feedback mechanism
Feedback control system
Fasting level / Normal value
(Modifier) level
After a meal

Table A2: Common multi-word expressions identified in the GLP-1 physiology corpus. These multi-word expressions represent self-contained triples and were consolidated for simplicity, with the understanding that they can be decomposed into their constituent terms during post-processing. In cases where the surface representation of a phrase was disjoint, annotators were instructed to combine the constituent terms into a single fragment phrase capturing the full expression.

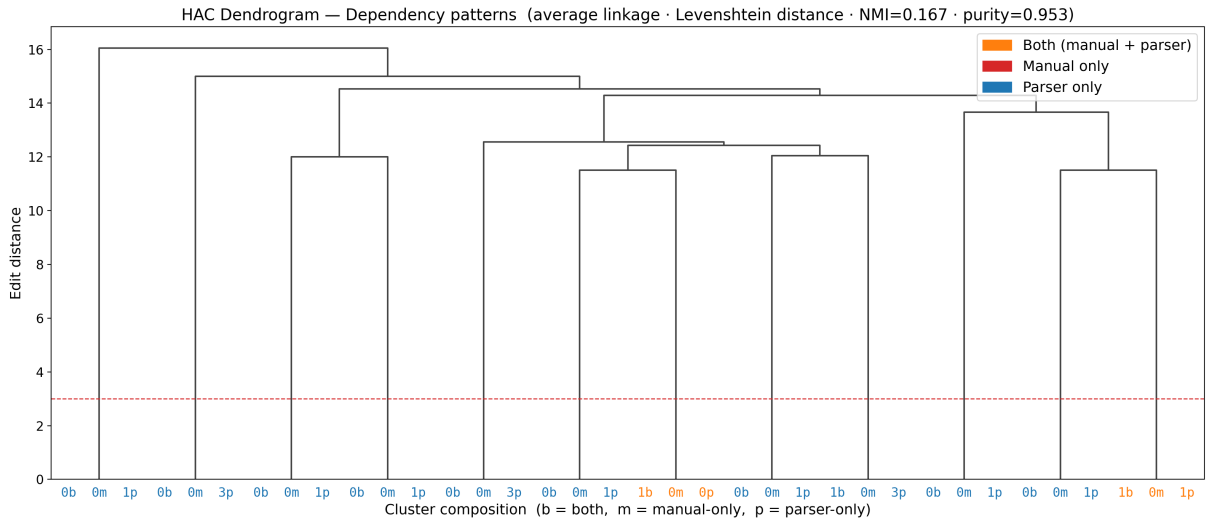


Figure A1: HAC dendrogram for dependency parser relation pattern clustering (average linkage, Levenshtein distance). The dashed red line indicates the cut height used to form final clusters; leaf labels show per cluster counts of patterns found in both sources (b), manual annotation only (m), and parser only (p). Clusters span edit distances up to 16, reflecting considerable structural diversity among dependency relation patterns.

Ontology	**Entities** (N = 29)	**Physiological Process** (N = 23)	**Physiological State** (N = 13)	**Relationship** (N = 55)
OLS ontologies				
OMIT	**0.52**	(0.13)	0.31	–
MESH	0.48	0.13	0.31	–
NCIT	0.38	**0.17**	**0.38**	**0.24**
CHEBI	0.24	–	–	–
HCAO	0.21	(0.04)	–	–
SNOMED	(0.17)	0.13	(0.31)	0.16
EFO	(0.14)	(0.09)	0.31	(0.02)
PRIDE	(0.07)	**0.17**	**0.38**	(0.05)
INO	–	**0.17**	–	(0.05)
BIOLINK	(0.03)	(0.09)	–	0.09
SLSO	–	–	–	0.07
AFO	–	(0.04)	(0.15)	0.07
Unmatched (OLS)	7	11	7	29
UMLS				
Coverage	0.55	0.35	0.23	0.11
Unmatched	13	15	10	49

Table A3: Ontology coverage of terminology categories (Entities, Physiological Process, Physiological State, and Relationship) queried against OLS and the UMLS. Coverage values reflect the proportion of distinct terms matched within each ontology. Coverage counts that are inside parenthesis are not part of the top five ontologies for that category, but still contain some coverage. (—) indicates no coverage was found. Unmatched counts reflect terms not found in any OLS ontology or the UMLS. OMIT=Ontology for miRNA Targets (Huang et al., 2011); MESH=Medical Subject Headings (Sewell, 1964); NCIT=National Cancer Institute Thesaurus (Sioutos et al., 2007); CHEBI=Chemical Entities of Biological Interest (Degtyarenko et al., 2007); HCAO=Human Cell Atlas Ontology (Welter et al., 2018); SNOMED=Systematized Nomenclature of Medicine—Clinical Terms (U.S. National Library of Medicine, 2026); EFO=Experimental Factor Ontology (Malone et al., 2010); PRIDE=Proteomics Identification Database Ontology (Perez-Riverol et al., 2022); INO= Interaction Network Ontology (Özgür et al., 2016); BIOLINK=Biolink Model (Unni et al., 2022); SLSO=Space Life Sciences Ontology (Berrios et al., 2024); AFO=Allotrope Foundation Ontology (Millecam et al., 2021); UMLS=Unified Medical Language System (Bodenreider, 2004)

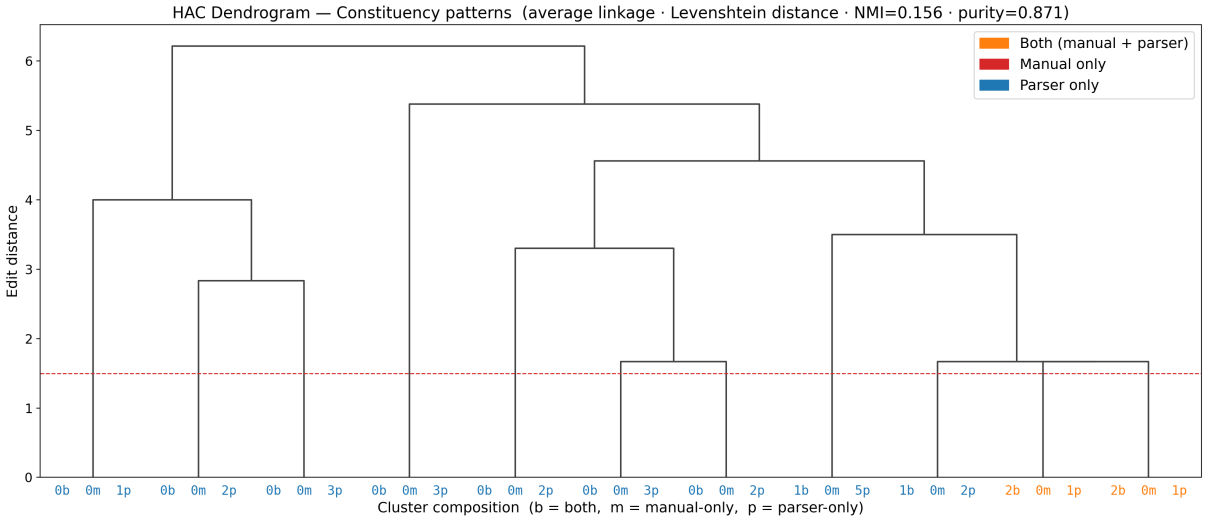


Figure A2: HAC dendrogram for constituency parser relation pattern clustering (average linkage, Levenshtein distance). he dashed red line indicates the cut height used to form final clusters; leaf labels show per cluster counts of patterns found in both sources (b), manual annotation only (m), and parser only (p). Cluster span edit distances up to 6, considerably lower than dependency relation patterns, likely reflecting the constrained phrase-label vocabulary from which the constituency relation patterns are constructed.

CENT: Context Engineering Framework for Normalization of Clinical Trial Procedures

Sanya B. Taneja[1], Ziqing Ji[2], Hans Verstraete[1], Ali Samadani[1]

[1]Johnson and Johnson Innovative Medicine, [2]University of Washington

Correspondence: asamadan@its.jnj.com

Abstract

Clinical Concept Normalization is essential for clinical research applications involving trial protocols, such as patient-trial matching. Existing approaches focus heavily on specific domains and need large, annotated datasets. To address these challenges, we propose CENT, a context engineering framework that combines semantic matching for candidate retrieval and Large Language Model (LLM) prompting for disambiguation. CENT achieves superior performance on clinical procedures normalization in both binary and hierarchical metrics compared to semantic matching or LLM-only approaches, without requiring fine-tuning. Advanced prompt strategies, including Chain-of-Thought and Tree-of-Thoughts, achieve high performance at practical cost. CENT is scalable and adaptable for normalization across diverse clinical vocabularies in real-world clinical applications.

1 Introduction

Clinical Concept Normalization (CCN), often referred to as entity linking, standardizes terms in clinical texts, such as clinical trial protocols, to standard concepts in clinical coding systems (e.g., ICD, SNOMED-CT, LOINC). In practice, normalization for clinical research applications is often performed as an ad hoc step, either as a manual effort or more commonly with string matching before downstream tasks, such as patient-trial matching (Wong et al., 2025). This limits both scalability and reliability in real-world deployments, where concept variations and ambiguous terminology are common (Wong et al., 2023).

For clinical trial protocols, CCN ensures reliable information retrieval and enables comparative analysis across trials in terms of eligibility criteria, disease evaluation steps or schedule of activities, predictive modeling using data from historic trials, and patient trial matching (Danhauser et al., 2025; Lopez et al., 2025).

Despite advances in semantic matching with pretrained and fine-tuned BERT-based models and, more recently, transformer models, biomedical entity linking remains challenging without domain-specific adaptation. Prompting general-purpose Large Language Models (LLMs) to assign codes to diagnosis and procedure descriptions has shown subpar performance and hallucinated codes (Soroush et al., 2024). Moreover, relying solely on LLMs for entity linking requires fine-tuning for each clinical coding system, which is expensive (Berkowitz et al., 2025).

We introduce **C**ontext **E**ngineered **N**ormalization for Clinical **T**rials (CENT) framework that leverages semantic matching to retrieve highly relevant information, which is then used in structured prompts to perform CCN. In CENT, semantic matching efficiently generates high-recall candidate matches, while LLM prompting improves disambiguation among closely related candidates without the need for extensive annotated data or vocabulary-specific fine-tuning.

Objective: To develop and evaluate a context engineering framework, CENT, for CCN in trial protocols, focusing on laboratory procedures, and benchmarking performance using binary and hierarchical metrics on a publicly available dataset.

Contributions:

- A context engineering framework, CENT, that combines semantic matching for context retrieval with LLM prompting for CCN.
- Evaluation of CENT using various prompt strategies on a publicly available clinical procedures dataset, assessed with both binary and hierarchical metrics.
- Comparative evaluation of CENT against embedding models and LLM-only prompting for CCN.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 729–741
July 3-4, 2026 ©2026 Association for Computational Linguistics

1.1 Related Work

Current state-of-the-art CCN systems and benchmark datasets primarily focus on normalizing diseases, symptoms, and chemicals, while clinical procedures and laboratory measurements are underrepresented. This results in incomplete normalization of clinical trial protocols and is a significant barrier for real-world automated protocol analysis and benchmarking. While CCN tools such as MedLinker, SciSpacy, and cTAKES support entity linking for procedures (Kartchner et al., 2023), these systems typically rely on string matching, leading to suboptimal performance.

Semantic matching models, such as SapBERT (Liu et al., 2021) that is fine-tuned with Unified Medical Language System (UMLS) concepts, have demonstrated strong similarity-based matching for biomedical entities (Kartchner et al., 2023). Semantic matching approaches with domain-specific embeddings achieve high recall and are effective at candidate generation, however, they struggle in candidate disambiguation as the size of the target vocabulary increases (Kartchner et al., 2023). Moreover, these systems are primarily optimized for diseases, drugs, and chemicals, as most benchmark datasets (e.g., NCBI Disease, BC5CDR) focus on these domains (Ferré et al., 2020; Yang et al., 2023).

Recent work has explored the use of LLMs and generative AI approaches for CCN. Berkowitz et al. (2025) developed RAGNorm, a system that combines semantic matching with LLM prompting to normalize biomedical texts to SNOMED-CT codes. RAGNorm uses general-purpose models for both semantic matching and prompting, and while it outperformed the baseline semantic matching methods, the authors highlight challenges with generalizability and a decline in performance with increased target vocabulary size. Soroush et al. (2024) found that prompt-based LLMs frequently hallucinate invalid or non-specific codes when relying on internal knowledge of clinical vocabularies, and that prompt engineering alone could not consistently produce correct codes for all concept descriptions. Other studies have noted the high computational cost and limited improvements with LLM-only pipelines compared to semantic matching and domain-specific embeddings in the biomedical normalization task (Dobbins, 2024; Wang et al., 2024). Jahan et al. (2024) observed that baseline models such as BioBART still outperform LLMs in recall for chemical and disease normalization, especially with small datasets and large vocabularies.

Combining the strengths of semantic matching and LLM prompting, however, appears to be a promising approach for CCN in specialized domains (Borchert et al., 2024; Berkowitz et al., 2025). Strategic prompt design with curated context beyond simple prompts can help overcome the limitations of LLM-only normalization approaches. In this work, we present CENT, a framework that integrates 1) domain-specific semantic matching to retrieve relevant context, followed by 2) a structured prompt design that embeds the retrieved context, and 3) explicit evaluation rubric to guide the LLM in performing CCN.

Evaluating CCN systems remains challenging due to variability across datasets, biases in scoring metrics, the lack of clear separation between entity recognition and subsequent normalization tasks, and the predominant focus of benchmark datasets on diseases and drugs (Kartchner et al., 2023; Ferré et al., 2020). Furthermore, prior evaluations rely primarily on binary metrics such as accuracy, precision, and recall, overlooking the hierarchical nature of clinical coding systems, which warrants more nuanced assessments that account for relationships among concepts (Ferré and Langlais, 2023). To address these gaps, in this study, we adapt hierarchical evaluation metrics to provide a more comprehensive assessment of CCN performance for normalization of clinical procedures.

2 Datasets

In this work, we evaluate our proposed CCN framework on the LOINC to CPT mappings of laboratory procedures available through UMLS (National Library of Medicine, 2006). LOINC, created by the Regenstrief Institute, is the universal standard for laboratory tests and clinical observations (Regenstrief Institute, 2025). The CPT vocabulary, created and maintained by the American Medical Association, is used for reporting procedural activities and services for billing and reimbursement (American Medical Association, 2025). Both vocabularies are commonly used for coding of activities related to clinical trials.

The UMLS LOINC–CPT mappings dataset contains 2,019 LOINC concepts linked to CPT codes; in some cases, multiple LOINC concepts map to a single CPT code. We excluded mappings to 21

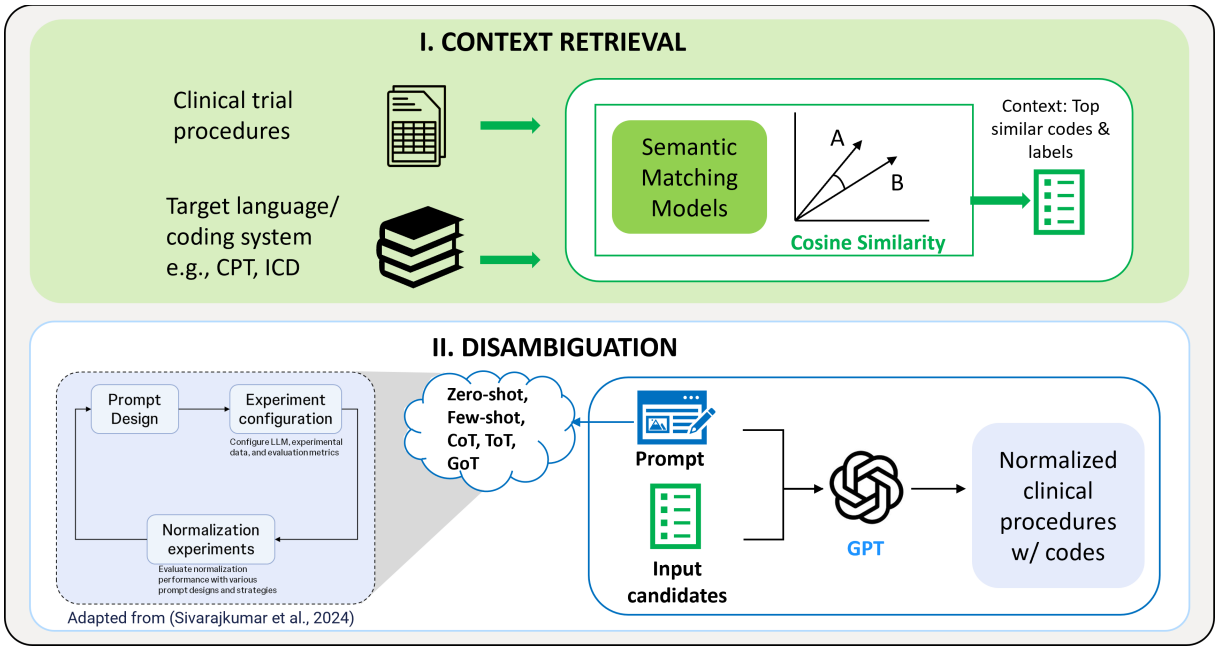


Figure 1: Overview of the CENT framework (I) Context retrieval with semantic matching, and (II) context engineering and prompt designs for disambiguation with GPT models. Iterative prompt design is based on prior work (Sivarajkumar et al., 2024) with zero-shot, few-shot, Chain-of-Thought (CoT), Tree-of-Thoughts (ToT), and Graph-of-Thought (GoT) prompt strategies.

deprecated CPT codes (44 mappings removed). We also excluded three non-specific allergen-related CPT codes—86001, 86003, and 86005—removing 314, 1,017, and 95 mappings, respectively, as these codes are unrelated to trial procedures and outside the scope of normalization. After these exclusions, 593 LOINC-CPT mappings remained. LOINC and CPT concept labels were further processed to remove punctuation and tokenized (average LOINC tokens = 7.6, average CPT tokens = 11.3).

We used CPT vocabulary (v2015) codes and labels as the target concept data for evaluation, as this was the last publicly available version of CPT. These CPT concepts were filtered to include the measurement domain only, resulting in 1,584 CPT target codes and labels.

2.1 Protocol-derived Datasets

We additionally applied CENT to two internal datasets of clinical assessments extracted from schedule of activities tables in clinical trial protocols. Dataset 1 includes 669 unique laboratory, diagnostic, survey, and billing procedures partially mapped to CPT codes, and Dataset 2 includes 139 laboratory/biomarker procedures from 20 trials mapped to a custom laboratory vocabulary (N=248 concepts). We applied CENT to predict target concepts for these datasets. For these datasets, we report quantitative metrics and expert review outcomes for the experiments with these protocol-derived datasets.

3 Methodology

We propose the CENT framework for CCN that systematically combines semantic matching with LLM prompting. CENT is designed to retrieve highly relevant context as top-ranked candidate codes and labels, and deliver it to the LLM in the form of structured, sectioned prompts (Figure 1).

3.1 Context Retrieval with Semantic Matching

In the first stage, CENT uses semantic matching with two fine-tuned domain-specific BERT-based models, SapBERT and BioBERT-mnli, to retrieve candidate concepts from the target vocabulary. SapBERT is a word-based model fine-tuned with UMLS concepts for medical entity linking and generates candidates based on closeness of concept embeddings in semantic space (Liu et al., 2021). BioBERT-mnli is a sentence transformer model fine-tuned with natural language inference and text similarity datasets that has been used to effectively classify sentences from randomized controlled trials (Deka et al., 2022). For each input term (e.g., LOINC concept), the models embed the concept

label and the resulting embedding is used to compute cosine similarity scores with embeddings of the target concepts (e.g., CPT concepts). The top 5 concepts with the highest scores from each model are added to a set and the scores are normalized. The concepts are then sorted by similarity scores and de-duplicated. The top unique candidates (up to 10) are used as the curated context formatted as code + label. Note that similarity scores are not passed to the LLM. By deliberately constraining the search space, this step serves as context engineering, shaping and refining the information available for downstream reasoning and matching.

Implementation: For SapBERT, we tokenize the texts with truncation enabled, embed in batches, with padding of all inputs to maximum tokens within the batch. For BioBERT-mnli, we use sentence-transformers with mean pooling, truncation, and dynamic batch padding. Both models generate 768-dimensional embeddings. We compute cosine similarity between input and target embeddings and retrieve the top k candidates per model. To further characterize the retrieval stage, we report recall@k on the LOINC-CPT dataset, where a query is counted as retrieved if ground truth CPT appears in the top-k candidate set.

3.2 Structured Prompts for Disambiguation

In the second stage, the query code and its label along side with the curated contextual information from the previous step are embedded in a structured prompt for the LLM. Prompts are organized into clearly defined sections, including 1) description of the CCN task, 2) explicit rubric to guide the LLM in performing CCN, 3) instructions to score LLM thoughts, 4) examples of the CCN to be done along with reasoning for the selected matches, and 5) the input query and the list of candidate matched codes with their labels (curated context). The LLM is instructed to select the single best candidate or return "no match" if appropriate.

For the CCN task in this paper, we designed zero-shot and few-shot prompt strategies using the DSPy library (Khattab et al., 2023) and adapted the open-source Graph-of-Thoughts repository (Besta et al., 2024) to implement the Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thoughts (ToT) (Yao et al., 2023), and Graph-of-Thoughts (GoT) strategies. Broadly, the operations involved are (1) prompt the LLM (*generate_prompt*) to map input concept to corresponding target concept based on the context of candidate codes and labels, (2) parse and score the LLM response (*score_prompt*), (3) improve the final response based on the context (*tot_improve_prompt*) or aggregate responses (*aggregate_prompt*), and (4) keep the best scoring response. The scores are calculated by the LLM on a scale of 0-5 using the *score_prompt*. For details of prompts used in the task, see Appendix A. Each prompt strategy was tested and improved iteratively. The prompt design and iterative evaluation was based on prior research on task-specific prompt design and improved performance of heuristic and ensemble prompts in complex clinical information extraction and disambiguation tasks (Sivarajkumar et al., 2024).

Implementation: ToT is defined with branch factor 10 and depth 3 while GoT is defined with branch factors 10 and 20. We used the GPT4.1 (OpenAI, 2025a) and GPT4o-mini (OpenAI, 2024) models to assess the impact of the model size on performance and cost with temperature equal to 1.0 and maximum output tokens equal to 10000.

For the comparative evaluation with direct LLM prompting without engineering context, we provided the input term along with 1,584 CPT target codes and labels to the GPT4o-mini model. The total prompt and context was approximately 65,000 tokens (context window of model is 100,000 tokens). For comparative context-retrieval models, we utilized the OpenAI API for text-embedding-ada-002 model (OpenAI, 2022). We accessed GPT models via an Azure OpenAI endpoint with access-controlled API keys. CENT is designed to be model-agnostic and can be used with private endpoints and locally deployed LLMs, if required.

4 Evaluation

We evaluated CENT with binary metrics, hierarchical metrics, and total cost (in US dollars) for CCN on the LOINC-CPT dataset. Binary metrics included precision@5, recall@5, and F1 score. Cost was measured by the input and output tokens in the prompts based on the OpenAI API pricing for the models (OpenAI, 2025b). We report the total cost of the LOINC-CPT normalization task with 593 mappings. The performance evaluation also included comparison against: 1) semantic matching only using SapBERT, 2) semantic matching only using OpenAI text-embedding-ada-002 model, 3) context retrieval using OpenAI text-embedding-ada-002 model followed by LLM prompting for disambiguation, and 4) direct LLM prompting without

engineered context.

4.1 Hierarchical Metrics

Hierarchical metrics take into account hierarchical characteristics of predicted codes in CCN, including concept position within the ontology, ancestral and descendant information, and relationship with the correct match. Unlike binary metrics that only determine if a predicted code is an exact match, hierarchical metrics incorporate partial correctness, capture ontological distance, and reward higher semantic similarity for close mismatches. We included two hierarchical metrics in our evaluation, Wu-Palmer similarity and Resnik similarity.

Wu-Palmer Similarity. Wu-Palmer Similarity (Wu and Palmer, 1994) is a structural similarity measure that quantifies the closeness of two concepts within an ontology, based on their depth (distance from the root concept) and the depth of the lowest common ancestor (LCA). For concepts a and b:

$$\text{Wu–Palmer}(a,b) = \frac{2 \times \text{depth}(\text{LCA}(a,b))}{\text{depth}(a) + \text{depth}(b)} \quad (1)$$

The score ranges from 0 to 1, with higher values indicating greater semantic overlap.

Resnik Similarity. Resnik similarity (Resnik, 1995) is calculated as the information content (IC) of the LCA of two nodes in the hierarchy to measure the closeness of two nodes by how informative their LCA is.

$$\text{Resnik}(a,b) = IC(\text{LCA}(a,b)) \quad (2)$$

IC is calculated using the proposed method by Seco et al. (2004) to use the ontology's structure:

$$ic(c) = 1 - \frac{\log(\text{descendants}(c) + 1)}{\log(\text{total concepts})} \quad (3)$$

Resnik similarity is higher for concepts with LCA lower in the hierarchy.

The UMLS API (National Library of Medicine, 2025) was used to extract ontological relationships, including ancestors, descendants, and depth for CPT codes. Ancestors and descendants were identified recursively through the UMLS API endpoints *parent* and *children* to preserve the hierarchical order. For cases with "no match", hierarchical metrics values were set to 0.

5 Results and Discussion

Table 1 presents the evaluation results of CENT on the LOINC-CPT dataset. CENT consistently outperformed the comparative methods, including semantic matching alone and direct LLM prompting-across both binary and hierarchical metrics. The highest F1-score was achieved by CENT with CoT prompting (0.82), while the highest Wu-Palmer similarity (0.86) and Resnik similarity (0.87) were observed with ToT prompting. When compared to embedding-only approaches, CENT achieved a marked improvement in precision (highest precision 0.86 with CENT vs 0.15 with SapBERT-only) and Wu-Palmer similarity (0.77-0.86 with CENT vs 0.53 with SapBERT-only), indicating that normalized codes by CENT are not only more accurate but, if not an exact match, are closer to the correct matches in the ontology. Direct LLM prompting performed poorly across all metrics with prohibitive costs, confirming prior findings that this approach is infeasible for CCN (Soroush et al., 2024).

A comparison of embedding models within CENT (SapBERT + BioBERT-mnli) and a general-purpose OpenAI model (text-embedding-ada-002) further underscores the value of domain-specific embeddings for CCN. Using zero-shot prompting, CENT with SapBERT + BioBERT-mnli achieved an F1 score of 0.79, compared to 0.64 for the OpenAI text-embedding-ada-002 model followed by LLM zero-shot prompting. These findings corroborate prior reports that domain-specific models outperform general-purpose models on clinical data classification tasks (Gao et al., 2025).

Across prompt strategies, CENT with CoT achieved high performance (F1-score 0.82 with GPT4.1 and 0.80 with GPT4o-mini), while ToT and GoT were stronger in hierarchical metrics (Wu-Palmer similarity 0.86 and 0.85, respectively). GoT, despite adding operational complexity and cost, did not offer improved performance except for the lowest percentage of "no match" predictions (0%), followed by ToT (0.3%). The "no match" percentage indicates the failure to identify any matches. In contrast, CENT with zero-shot and few-shot prompts resulted in the highest percentage of "no match" predictions (14.1% and 12.8%, respectively). Error analysis has shown that, when allowed, CENT over-predicts codes (i.e., predicted code set includes more than one candidate) rather than "no match".

Cost is a critical factor for real-world and re-

Approach	Prompt Strategy	Precision @5	Recall @5	F1	Wu-Palmer	Resnik	% "no match"
CENT GPT4.1	Zero-shot	**0.87**	0.74	0.80	0.77	0.77	14.8
CENT GPT4.1	Few-shot	0.86	0.76	0.81	0.79	0.79	10.8
CENT GPT4.1	CoT	0.86	0.79	**0.82**	0.84	0.84	5.1
CENT GPT4.1	ToT	0.81	**0.81**	0.81	**0.86**	**0.87**	0.3
CENT GPT4.1	GoT	0.79	0.78	0.78	0.85	0.85	0.7
CENT GPT4o-mini	CoT	0.82	0.78	0.80	0.84	0.85	7.2
CENT GPT4o-mini	ToT	0.79	0.79	0.79	0.85	0.86	0.5
CENT GPT4o-mini	GoT	0.78	0.78	0.78	0.85	0.86	**0**
Embedding only (SapBERT)	-	0.15	0.74	0.25	0.53	0.56	-
Embedding only (ada-002)	-	0.15	0.76	0.25	0.55	0.62	-
ada-002 + GPT4.1	Zero-shot	0.64	0.64	0.64	0.81	0.82	0.3
LLM only (GPT4o-mini)	Few-shot	0.05	0.17	0.08	0.50	0.58	3.1

Table 1: **Top.** Evaluation of CENT with binary and hierarchical metrics for LOINC-CPT mappings (N=593). "no match" indicates cases where the model did not return any normalized code. **Bottom.** Evaluation of comparative approaches: (i) SapBERT only for CCN, (ii) OpenAI ada-002 for context retrieval and GPT4.1 for disambiguation, (iii) OpenAI ada-002 only, and (iv) LLM-only with GPT4o-mini for CCN.

Approach	Prompt Strategy	F1-score	Cost (USD)
CENT*	Zero-shot	0.79	0.36
CENT*	Few-shot	0.80	0.37
CENT*	CoT	**0.82**	12.89
CENT*	ToT	0.81	37.57
CENT*	GoT	0.78	59.89
CENT$^+$	CoT	0.80	0.56
CENT$^+$	ToT	0.79	2.66
CENT$^+$	GoT	0.78	3.27
LLM-only$^+$	Few-shot	0.08	102.5

Table 2: Cost (in USD) of API calls and F1-score for the prompt strategies used to evaluate CENT with LOINC-CPT dataset (N=593). "*" denotes GPT4.1 used for disambiguation and "$^+$" denotes GPT4o-mini model.

search applications. Table 2 presents the cost (in USD) of the API calls used for CENT with different prompt strategies in the LOINC-CPT normalization. While ToT and GoT prompting have higher costs ($37.57 and $59.89, respectively) compared to other strategies, our experiments showed that comparable performance is achievable with the smaller GPT4o-mini model using these advanced prompt strategies, but with significantly lower cost (e.g., cost for CENT with GPT4o-mini & ToT was $2.66 vs $37.57 for CENT with GPT4.1 & ToT).

Table 3 presents the results of the retrieval ablation analysis in the LOINC-CPT mapping experiment. Union of candidates from both models (SapBERT and BioBERT-mnli) achieves the highest recall (recall@10 improves to 0.85), improving the downstream CCN task driven by the retrieved context (union of candidates). Moreover, increasing the number of per-model candidates (k) from 5 to 10 produces no appreciable gains, indicating that CENT remains effective even with a small number of candidates.

To the best of our knowledge, this work represents the first implementation of advanced prompt strategies (ToT, GoT) for the CCN task. Compared to general tasks evaluated by Besta et al. (2024) where GoT outperformed other strategies, it is evident that model and prompt strategy selection for domain-specific tasks such as CCN cannot be gen-

Retrieval	Recall @1	Recall @5	Recall @10
SapBERT only	0.51	0.74	0.80
BioBERT-mnli only	0.59	0.79	0.82
Union (SapBERT ∪ BioBERT-mnli)	–	0.83	0.85

End-to-end (CENT+CoT)	F1	% "no match"
Union candidates, $k = 5$ per encoder	0.82	5.1
Union candidates, $k = 10$ per encoder	0.83	5.3

Table 3: Retrieval ablation and end-to-end sensitivity for LOINC-CPT mappings. Recall@k measures whether the ground truth CPT code is present in the top-k retrieved candidates.

eralized out of the box. The combined use of binary and hierarchical metrics provides a more complete evaluation: binary metrics alone may over- or underestimate model performance, but hierarchical metrics can indicate whether the predicted codes are proximally closer to the ground truth code in the ontology. For instance, CENT with ToT prompting showed lower F1-scores than CoT but higher Wu-Palmer and Resnik scores, indicating that even when predictions are incorrect, they are semantically closer to the ground truth.

Internal Validation: We applied CENT to the protocol-derived procedures, which are noisier and less standardized than LOINC labels. Examples of procedures included laboratory panels (chemistry, hematology), abbreviated texts (PSA, UPEP), imaging procedures (MRI Abdomen), biomarker assays (Serum biomarkers). For the 139 laboratory/biomarker procedures normalized to custom vocabulary concepts (Dataset 2), CENT (few-shot) achieved precision@5=0.77, recall@5=0.65, and F1=0.71. This dataset included instances with multiple target concepts and instances with no valid matches; "no match" is counted correct only when the ground truth is empty. For the dataset with 669 procedures (Dataset 1), we ran CENT to predict CPT codes, and 3 clinician-experts manually reviewed a random sample of 59 laboratory procedures. Each of 3 experts reviewed 20 samples with partial overlap. The outcomes were 28/59 (47.5%) exact matches to expert annotation, 17/59 (28.8%) partial matches (sibling/parent relationships), and 14/59 (23.7%) incorrect CENT match when compared to expert annotations.

Together with the benchmark dataset, these results suggest CENT can operate on noisy and often non-standard clinical texts common in real-world data (e.g., clinical protocols). Notably, the optimal choice of models and prompt strategy should be guided not only by performance metrics but also by specific use case requirements, the target vocabulary, acceptable operational costs, and the prevalence of "no match" scenarios in the real-world data. CENT enables high-recall candidate generation and robust candidate disambiguation for CCN without model fine-tuning or annotated training datasets. CENT does not require model training and allows for substitution of prompt formats, LLMs, and reference clinical vocabulary system as needed for domain adaptation.

6 Conclusion

We present a context engineering framework, CENT, for CCN that combines semantic matching with domain-specific models for context engineering and LLM prompting for disambiguation. CENT consistently outperformed other methods - including semantic matching alone and LLM-only - on laboratory procedures with both binary and hierarchical metrics without any fine-tuning or annotated training data. CENT enables scalable, cost-effective normalization to support real-world clinical research applications. Ongoing work involves applying CENT to use cases in clinical trial optimization and benchmarking, and exploring further prompt strategies to refine normalization.

Limitations

This study has several limitations. First, the evaluation mainly focused on a single dataset with laboratory procedures, along with two internal datasets for validation. The framework will need to be evaluated on larger, diverse datasets with more entity types and clinical vocabularies. Second, prompt design is currently a manual process in CENT that must be tailored for other CCN tasks. Finally, our prompts were limited to GPT models and cost calculations are based on current API pricing. The performance may not reflect future changes in models and re-evaluation of CENT is warranted with newer LLMs.

References

American Medical Association. 2025. Current procedural terminology (cpt®).

Jacob S. Berkowitz, Apoorva Srinivasan, Jose Miguel Acitores Cortina, Yasaman Fatapour, and Nicholas P Tatonetti. 2025. Biomedical text normalization through generative modeling. *Journal of Biomedical Informatics*, 167:104850.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Florian Borchert, Ignacio Llorca, and Matthieu-P Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database: The Journal of Biological Databases and Curation*, 2024:baae067.

Katharina Danhauser, Yingding Wang, Christoph Klein, Uta Tacke, Larissa Mantoan, Laura Aurica Ritter, Florian Heinen, Chiara Nobile, and Moritz Tacke. 2025. Using large language models to extract information from pediatric clinical reports. *PLOS Digital Health*, 4(7):e0000919.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2022. Evidence extraction to validate medical claims in fake news detection. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, page 3–15, Berlin, Heidelberg. Springer-Verlag.

Nicholas J. Dobbins. 2024. Generalizable and Scalable Multistage Biomedical Concept Normalization Leveraging Large Language Models. *arXiv preprint*. ArXiv:2405.15122 [cs].

Arnaud Ferré, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. 2020. C-Norm: a neural approach to few-shot entity normalization. *BMC Bioinformatics*, 21(S23):579.

Arnaud Ferré and Philippe Langlais. 2023. An analysis of entity normalization evaluation biases in specialized domains. *BMC Bioinformatics*, 24(1):227.

Yuhe Gao, Runxue Bao, Yuelyu Ji, Yiming Sun, Chenxi Song, Jeffrey P Ferraro, and Ye Ye. 2025. Transfer learning with clinical concept embeddings from large language. *AMIA Summits on Translational Science Proceedings*, 2025:167.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S. Mitchell. 2023. A Comprehensive Evaluation of Biomedical Entity Linking Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2023:14462–14478.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P. Ma, April S. Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, Nigam H. Shah, and Jonathan H. Chen. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

National Library of Medicine. 2006. Loinc to cpt mapping. `https://www.nlm.nih.gov/research/umls/mapping_projects/loinc_to_cpt_map.html`. Accessed: 2024-02-03.

National Library of Medicine. 2025. Umls rest api documentation. `https://documentation.uts.nlm.nih.gov/rest/home.html`. Accessed: 2026-02-03.

OpenAI. 2022. text-embedding-ada-002.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2025a. GPT-4.1.

OpenAI. 2025b. OpenAI API Pricing.

Regenstrief Institute. 2025. Loinc: A universal standard for identifying laboratory observations.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

Nuno Seco, Tony Veale, and Jer Hayes. 2004. An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, page 1089–1090, NLD. IOS Press.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N. Nadkarni, and Eyal Klang. 2024. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI*. Publisher: Massachusetts Medical Society.

Andy Wang, Cong Liu, Jingye Yang, and Chunhua Weng. 2024. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association*, 31(9):2076–2083.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Cliff Wong, Sam Preston, Qianchu Liu, Zelalem Gero, Jaspreet Bagga, Sheng Zhang, Shrey Jain, Theodore Zhao, Yu Gu, Yanbo Xu, Sid Kiblawi, Srinivasan Yegnasubramanian, Taxiarchis Botsis, Marvin Borja, Luis M. Ahumada, Joseph C. Murray, Guo Hui Gan, Roshanthi Weerasinghe, Kristina Young, and 6 others. 2025. Universal Abstraction: Harnessing Frontier Models to Structure Real-World Data at Scale. *arXiv preprint*. ArXiv:2502.00943 [cs].

Cliff Wong, Sheng Zhang, Yu Gu, Christine Moung, Jacob Abel, Naoto Usuyama, Roshanthi Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. 2023. Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, pages 846–862. PMLR. ISSN: 2640-3498.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Siyu Yang, Peiliang Zhang, Chao Che, and Zhaoqian Zhong. 2023. B-LBConA: a medical entity disambiguation model based on Bio-LinkBERT and context-aware mechanism. *BMC Bioinformatics*, 24(1):97.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

A Appendix: Prompts

We present the prompts for the LOINC-CPT normalization task. **zero_shot** and **few_shot** prompts included simple instructions for the task along with the input LOINC concept and candidate concepts from context retrieval. For Chain-of-Thought (CoT), Tree-of-Thoughts (ToT), and Graph-of-Thought (GoT) prompting, we extended the graph-of-thoughts framework (Besta et al., 2024) to create a LOINC-CPT normalization task. Graph-of-thoughts describes prompt operations as 'transformations of thoughts' that ultimately give the final response from the language model. The **generate_prompt** generates new thoughts from a single thought with a branching factor to define the number of thoughts generated. The **tot_improve_prompt** enhances a given thought using information provided in another thought (could be generated previously). The **aggregate_prompt** aggregates thoughts into new ones in GoT. Thus, for the LOINC-CPT normalization task, CoT was modeled with the `generate` operation, ToT with `generate` and `improve` operations, and GoT with `generate` and `aggregate` operations. All three prompt strategies used the **score_prompt** for scoring the LLM responses. Examples for the prompts were generated from UMLS, however, these can be alternately created by experts or by varying the ground truth data. Other configurations of the operations can be implemented depending on the task and desired complexity of operations.

zero-shot prompt:

<Instruction>

You are an expert clinical assistant. You are provided with a LOINC measurement concept delimited by <> and a list of candidate CPT concepts for matching, delimited by []. These candidates are generated using a BERT model, which assesses the similarity between the embeddings of the LOINC query and all CPT concepts. The first candidate in the ranked list is the most similar CPT concept to the LOINC query based on BERT similarity. For the given LOINC concept and candidate list, your task is to -

1. Assess the candidate CPT concepts based on your knowledge about laboratory terms and clinical vocabularies. Create a new candidates list by reordering and/or removing concepts. If the original list is correct, do not create a new list. If there are no good candidate concepts for the LOINC concept, create an empty list.
2. Then using the new list of candidates, select the best match for the given LOINC concept. There is only one best CPT match in the candidate list for the given LOINC concept. A good match in the candidate list of CPT concepts is when the candidate exactly matches the LOINC concept or is related to the LOINC concept.
3. Return the matched code in a list selected only from the candidates provided. If there is no good match in the list of candidates, return ['no match']. Return the output as list only and no other text.

</Instruction>

Term: <term>
Candidates: [candidates]

---

few-shot prompt:

<Instruction>

You are an expert clinical assistant. You are provided with a LOINC measurement concept delimited by <> and a list of candidate CPT concepts for matching, delimited by []. These candidates are generated using a BERT model, which assesses the similarity between the embeddings of the LOINC query and all CPT concepts. The first candidate in the ranked list is the most similar CPT concept to the LOINC query based on BERT similarity. For the given LOINC concept and candidate list, your task is to -

1. Assess the candidate CPT concepts based on your knowledge about laboratory terms and clinical vocabularies. Create a new candidates list by reordering and/or removing concepts. If the original list is correct, do not create a new list. If there are no good candidate concepts for the LOINC concept, create an empty list.
2. Then using the new list of candidates, select the best match for the given LOINC concept. There is only one best CPT match in the candidate list for the given LOINC concept. A good match in the candidate list of CPT concepts is when the candidate exactly matches the LOINC concept or is related to the LOINC concept.
3. Return the matched code in a list selected only from the candidates provided. If there is no good match in the list of candidates, return ['no match']. Return the output as list only and no other text.

</Instruction>

<Examples>

Term: 22360-2:HTLV I IgG Ab [Presence] in Serum
Candidates: [("86687","Antibody HTLV-I"), ("86296","HEPATITIS A ANTIBODY HAAB IGG AND IGM"), ("86709","Hepatitis A antibody HAAb IgM antibody"), ("82784","Gammaglobulin immunoglobulin IgA IgD IgG IgM each"), ("86701","Antibody HIV-1"), ("81381","HLA Class I typing high resolution ie alleles or allele groups one allele or allele group eg B*5701P each"),

("86001","Allergen specific IgG quantitative or semi-quantitative each allergen")]
Answer: ["86687"]

Term: 14416-2:Calcium [Moles/volume] in Pleural fluid
Candidates: [("82330","Calcium ionized"), ("82310","Calcium total"), ("82331","Calcium after calcium infusion test"), ("82340","Calcium urine quantitative timed specimen"), ("82360","Calculus quantitative analysis chemical"), ("80410","Calcitonin stimulation panel eg calcium pentagastrin This panel must include the following Calcitonin 82308 x 3")]
Answer: ["82310"]
</Examples>

Term: <term>
Candidates: [candidates]

---

generate_prompt for CoT, ToT, GoT:

<Instruction>
You are an AI medical coding assistant designed to help healthcare providers accurately map LOINC codes to their corresponding or nearest CPT codes. You will be provided with a LOINC laboratory code and name in the format query_lab_code:query_lab_name, referred to as the LOINC query, hereafter. Occasionally, you may also receive a ranked list of candidate CPT code matches, with each match formatted as a list of tuples; for example: []. These candidates are generated using a BERT model, which assesses the similarity between the embeddings of the LOINC query and all CPT concepts. The first tuple in the ranked list is the most similar CPT concept to the LOINC query based on BERT similarity.

Your goal is to identify the most appropriate CPT match(es) for the LOINC query. A suitable CPT match should:

1. Exactly correspond to the LOINC concept, or
2. Be closely related (e.g., same lab family, sub-test of the LOINC query, or a broader lab panel that includes the LOINC query).

Your output **must** follow this exact format and be enclosed between <Match> and </Match> tags: <Match>[("CPT_CODE","CPT_LABEL"), ("CPT_CODE","CPT_LABEL")]</Match>

Example of correct output: <Match>[("86256","Fluorescent noninfectious agent antibody titer each antibody"), ("82330","Calcium ionized")]</Match>
</Instruction>

<Approach>
1. Assess Candidate List Availability: If a candidate list is provided, review each candidate, then reorder, add, or remove concepts as needed to refine the list. If no candidate list is provided (i.e., the list is empty, "None", or ""): Generate a list of potential CPT matches for the LOINC query.
2. Evaluate Candidates: Use the refined candidate list to select the best CPT match(es) for the LOINC query. There may be multiple suitable matches or none at all.
3. Return Results: Output a list of the best matches. If no matches are found, return ['no match'].
</Approach>

<Examples>
LOINC query: 22360-2:HTLV I IgG Ab [Presence] in Serum
Candidate matches: [("86687","Antibody HTLV-I"), ("86296","HEPATITIS A ANTIBODY HAAB IGG AND IGM"), ("86709","Hepatitis A antibody HAAb IgM antibody"), ("82784","Gammaglobulin immunoglobulin IgA IgD IgG IgM each"), ("86701","Antibody HIV-1"), ("81381","HLA Class I typing high resolution ie alleles or allele groups one allele or allele group eg B*5701P each"), ("86001","Allergen specific IgG quantitative or semi-quantitative each allergen")]
<Match>[("86687","Antibody HTLV-I")]</Match>

LOINC query: 14416-2:Calcium [Moles/volume] in Pleural fluid
Candidate matches: [("82330","Calcium ionized"), ("82310","Calcium total"), ("82331","Calcium after calcium infusion test"), ("82340","Calcium urine quantitative timed specimen"), ("82360","Calculus quantitative analysis chemical"), ("80410","Calcitonin stimulation panel eg calcium pentagastrin This panel must include the following Calcitonin 82308 x 3")]
<Match>[("82310","Calcium total")]</Match>
</Examples>

<Input>
LOINC query: {query_lab_code}:{query_lab_name}
Candidate matches: {candidates}
</Input>

---

tot_improve_prompt for ToT:

<Instruction>
The following list represents the list of CPT matches for the LOINC query query_lab_code:query_lab_name:cpt_matches. These CPT matches might include wrong matches or might be an empty list shown as []. Your task is to find the correct CPT matches. Make sure if there are identified CPT matches, they are truly related to the LOINC query by ensuring that these matches are either exactly correspond to the LOINC query, or are closely related. Your output must follow this exact format and be enclosed between <Match> and </Match> tags:

<Match>
["CPT_CODE": "<CPT_CODE>", "CPT_LABEL": "<CPT_LABEL>","CPT_CODE": "<CPT_CODE>", "CPT_LABEL": "<CPT_LABEL>"]
</Match>

Replace <CPT_CODE> with the actual CPT code match and <CPT_LABEL> with its corresponding label. Only return valid matches inside the tags — do not include extra text.
Example of correct output: <Match>[("86256","Fluorescent noninfectious agent antibody titer each antibody"), ("82330","Calcium ionized")]</Match>
</Instruction>

<Approach>
To fix the incorrect list of CPT matches:

1. If there are CPT matches:
   (a) For each pair of CPT match, check which one is more related to the LOINC query and retain it and drop the other.
   (b) Iterate over the pairs of CPT matches, until you are left with a single match.
   (c) If the remaining CPT match is the exact same concept as LOINC query, sibling concept, parent concept, keep it, otherwise, remove it from the list.
   (d) Return the resulting match if any
2. If there are no CPT matches, indicate as an empty list

You are an AI medical coding assistant designed to help healthcare providers accurately map LOINC codes to their corresponding or nearest CPT codes. Use your knowledge to find the best match for query_lab_code:query_lab_name.
</Approach>

<Examples>
LOINC query: 13051-8:Protein S [Units/volume] in Platelet poor plasma
Matches: [("85306","Protein S [Units/volume] in Platelet poor plasma"), ("85305","Clotting inhibitors or anticoagulants protein S total")]
Reason: 85306 is the exact match, while 85305 is a sibling match. <Match>[("85306","Protein S

[Units/volume] in Platelet poor plasma")]</Match>

LOINC query: 14252-1:Smooth muscle Ab [Presence] in Serum
Matches: [("86015","Actin smooth muscle antibody ASMA each"), ("86256","Fluorescent noninfectious agent antibody titer each antibody"),(86381,"Mitochondrial antibody eg M2 each")]
Reason: The actual match is 86256 which is not in the list of matches. The closest match 86255 is a sibling match, while 86015 and 86381 are cousin matches (more distant than the sibling match).
<Match>[("86256","Fluorescent noninfectious agent antibody titer, each antibody")]</Match>
</Examples>

LOINC query: {query_lab_code}:{query_lab_name}

---

score_prompt for CoT, ToT, GoT:

<Instruction>
The following is the list of CPT matches for LOINC query query_lab_code:query_lab_name: current. CPT matches are formatted as a list of tuples, where each tuple in the list represents a match and has only two elements, where the first element is a CPT code and the second element is its corresponding CPT label. Score each match in terms of how closely it represents the LOINC query. A score of 5 implies that the match is the exact match, a parent match gets a score of 4, a sibling match gets a score of 3, and grandparent or grand child matches get a score of 2, cousins get a score of 1, and everything else get a score of 0. You may provide reasoning for your scoring, but you should put the final scores in a list between the tags <Score> and </Score>. The order of final score list corresponds to the order of CPT matches i.e., the first and second scores correspond to the first and second CPT matches, respectively.
</Instruction>

<Example>
LOINC query: 13051-8:Protein S [Units/volume] in Platelet poor plasma
Matches: [("85306","Protein S [Units/volume] in Platelet poor plasma"), ("85305","Clotting inhibitors or anticoagulants protein S total")]
Reason: 85306 is the exact match, while 85305 is a sibling match.
<Score>[5,3]</Score>
</Example>
LOINC query: {query_lab_code}:{query_lab_name}

---

aggregate_prompt for GoT:

<Instruction>
The following list of lists presents the potential CPT matches for query_lab_code:query_lab_name: current. Combine and merge them into one list to only retain top 5 most relevant ones, with no repeated matches. Your output must follow this exact format and be enclosed between <Match> and </Match> tags:
<Match>[("<CPT_CODE>","<CPT_LABEL>"), ("<CPT_CODE>","<CPT_LABEL>")]</Match>
</Instruction>

Agentic AI Architectures for SOAP Note Generation

Keno Hanken
Independent Researcher
Berlin, Germany
keno-hanken@labs.arg-os.eu

Abstract

Clinical documentation places significant time demands on medical professionals, consumes institutional resources, and is prone to errors that may compromise patient care. Recent advances in LLMs offer promising approaches for automating clinical note generation; however, the impact of different AI architectural designs remains underexplored, particularly for agentic AI systems. This study compares three architectures — single-LLM, multi-agentic, and swarm-agentic — for automated SOAP (Subjective, Objective, Assessment, Plan) note generation from doctor–patient dialogues. All approaches employ QLoRA-finetuned Ministral 3 models (3B and 8B parameters) trained on the MedSynth dataset (Mianroodi et al., 2025), comprising 10,030 dialogue–note pairs across 2,006 ICD-10 code classes. Performance is evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore against a lexical-overlap baseline (dialogue vs. ground-truth SOAP, no inference). Results show that all finetuned models substantially outperform the baseline, while differences between architectural variants remain marginal. The single-LLM setup achieves the strongest performance across all metrics; 3B and 8B variants perform nearly identically on semantic similarity (BERTScore), while ROUGE differences are small but statistically significant. Qualitative inspection further reveals that residual differences across architectures are driven primarily by shared dataset priors rather than by architectural reasoning capacity. The results are based on synthetic data without human evaluation and reflect architectural behavior only.

1 Introduction

Clinical documentation serves multiple essential purposes including care continuity, research, and legal protection, but places significant time demands on medical professionals. The average consultant physician in clinical environments spends 12 hours a week adding to clinical documentation for patient records. This expenditure equals a productive monetary loss of 54,500 USD per year, per physician (Ignetica Ltd, 2022). Furthermore, 74% of medical professionals state that current efforts for clinical documentation impede patient care (American Medical Informatics Association, 2024). These findings suggest that a reduction of necessary efforts for patient records could address these and related challenges. However, a persisting challenge in documentation is the accurate and comprehensive documentation of relevant information. Medical professionals spend on average over an hour extra per day researching patient information (Ignetica Ltd, 2022). Adding to this, Bell et al. (2020) found in an empirical US study that 1 in 5 patient records contain at least one mistake, of which 40% are considered serious from the affected patients' perspective. Given these challenges, it is vital to make note-taking and information retrieval more time- and cost-effective.

Advances in LLM techniques offer a promising technical response to those challenges. There are countless strategies and process elements where AI can be used to enhance the process of note-taking, from transcribing speech, to reasoning about appropriate therapies for patients. Yet, the most impactful step is to ensure effective and accurate note composition from available data, such as transcripts. A widely adopted format for such notes is the SOAP schema (Subjective, Objective, Assessment, Plan), which structures clinical observations into four standardized dimensions and is detailed in Section 2.1. However, multiple strategies exist for how an AI infrastructure could solve the problem of composing clinical records, i.e. a single but powerful LLM could be employed to compose documentation, or a swarm of small, highly specialized AI agents could dissect the task of note-taking among them and collaboratively reason about the note composition. As for AI swarms, the landscape

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 742–752
July 3-4, 2026 ©2026 Association for Computational Linguistics

of empirical studies in the health-care domain is still an emerging field of research. This study aims to serve as an orientation for AI-architectural solutions for the task of clinical note-taking, answering the question of *"Do multi- / swarm-agent architectures outperform a single finetuned LLM for SOAP note generation?"*. The contributions of this paper are:

- A systematic evaluation of single-LLM, multi-agent, and swarm-agentic architectures for clinical SOAP note generation under controlled conditions.
- Highlighting that increased architectural complexity does not yield meaningful performance gains, questioning the practical utility of agentic approaches for this task.
- A reproducible experimental setup for SOAP note generation based on the MedSynth dataset, including consistent training and evaluation procedures.

The remainder of this paper is structured as follows: Section 2 shows the academic background, Section 3 describes the experimental setup and model architectures, Section 4 presents the results, Section 5 discusses the findings, and Section 6 concludes.

2 Literature

2.1 Theoretical Background

Weed (1964) proposed guidelines for the composition and structure of medical records. Each one of the 5 rules addresses a distinct viewpoint of the available facts and assumptions during a medical consultation. The framework prescribes (1) review and organization of all available data sources for a holistic patient view; (2) interpretation of those data in light of the clinician's impression of the patient; (3) derivation of a treatment plan through inductive reasoning over the contextualized data; (4) continuous data acquisition aligned with the evolving treatment plan; and (5) explicit articulation of the relationships between data entities (Weed, 1964). Through this, Weed defined core principles of data-driven patient care. His data organization framework laid the groundwork for the now widely adopted SOAP note schema.

Podder et al. (2023) lay out the current widely used adoption of the SOAP schema. A SOAP note is composed of 4 dimensions, where:

- S stands for *Subjective*, capturing what the patient reports themselves (chief complaint, history of present illness, etc.).
- O stands for *Objective* and captures data obtained through observation or empirical measurement (i.e. ECG, RR, observed skin condition, etc.).
- A stands for *Assessment* and concerns the analysis of the S and O findings to deductively state a diagnosis and/or clarify existing differential diagnoses.
- P stands for *Planning* and lays out how a patient's condition is treated methodically and / or further monitored and assisted. (Podder et al., 2023)

Bommarito et al. (2025) define the criteria for an agentic system, by comparing successive levels of systems that process and act on data. A level 1 agentic system is defined as *"[...] any entity that pursues goals through perception and action, with at least minimal discretion over which action to take in response to what it perceives."* (Bommarito et al., 2025) meaning a system that possesses a *goal* to achieve, a *perception* of the environment, and the *actionable capabilities* to affect their environment. An example could be a thermostat which maintains a certain level of temperature through temperature measurements. It is important to notice that a level 1 agent's properties are not limited to the described properties, but also can be extended by one or more of the 3 additional properties, that a level 2 adds. An agentic system is considered to be level 2 if the system's logic follows specific *rules* that are deterministic and non-flexible. It extends level 1 agent capabilities by possessing a (continuous) *iterative* data / environment observation, an *adaptation* to challenges and new observations, and a *termination* under defined criteria. An example could be a web scraper that monitors news websites for specified keywords and notifies subscribers. A level 3 agentic system however leverages AI architectures like neural networks, LLMs, or VLMs with traditional computing resources, while possessing the same properties as a level 2 agent. A level 3 agent, for example, can be a trading bot that dynamically adjusts strategies based on government announcements and news. Bommarito et al. clarify that the definition of an agent is bound to, but not limited to, a system's capabilities to influence its environment

in different levels with different methods. Furthermore, the authors advocate for a clear definition of the term *agent*, as the term itself does not reflect the different tasks and properties an intelligent system possesses (Bommarito et al., 2025).

2.2 Related Work

Li et al. (2025) introduce an expert curated simulated dataset called *CliniKnote* of 1,200 doctor–patient conversations with clinical notes. For these notes, the SOAP schema was extended by a keyword section on top, creating the K-SOAP schema, to enable quick identification of relevant symptoms and diseases. Furthermore, the authors benchmarked the K-SOAP note generation across 22 LLMs via metrics such as ROUGE, BERTscore, Bleurt (Sellam et al., 2020), and QuestEval (Scialom et al., 2021). For keyword extraction and generation, a combination of clinical named entity recognition and relation extraction for symptoms and diseases from dialogues has been used. The authors defined 8 entity types combining 5 relational prefixes with symptom and disease categories. If a conversational contribution was identified to be relevant for one of the specified categories, the information was tagged. The benchmark was conducted across all tested LLMs in the 3 categories: baseline models, open-source LLMs, and commercial models. For open-source and baseline models, the authors used QLoRA to reduce resource demands, training adapters for finetuning in three modes: a single adapter for the full note, 4 adapters for grouped sections, or 15 separate adapters for each individual section (denoted with a *section-4* / *section-15* suffix). Commercial models were instead evaluated via zero-shot and one-shot in-context learning. The results showed that the GPT-4o model in one-shot mode (zero-shot was tested as well for closed-source models) outperformed all tested open source models and baseline models marginally in 7 out of 9 metrics, compared to the next best performing model qCammel-13b-section-15, a model optimized for academic medical knowledge. Among the baseline and open-source LLMs, the model qCammel-13b-section-15 outperformed other models moderately in 7 out of 9 metrics, compared to the second best performing model Llama3-8b-section-4 (Li et al., 2025). These results suggest that the type and quality of domain adaptation matter more than model size - as evidenced by OpenBioLLM's poor performance despite its biomedical specialization. However, GPT-4o's superior one-shot results indicate that scale remains a contributing factor.

Ramprasad et al. (2023) analyzed an approach for generating SOAP notes based on conversational data. The authors dissected the SOAP dimensions beyond the 4 well-known dimensions, down to the contained scopes of information each dimension holds. For example, the S-dimension holds information about the chief-complaint, past medical history, allergies, etc. In total, the study features 12 subdimensions. In their approach, the authors compared the performance of a general-purpose model (GPT 3.5) in zero- and few-shot settings, against a default finetuned BART-model, and a BART-model with modified architecture. The authors also compared 3 architectural modifications of the BART-model to support the identification of SOAP-specific information from the conversation input. **BART+EMB** adds SOAP-(sub-)section tokens to the BART-embedding. It was finetuned with annotated and tagged data, which provides explicit conditioning on the target SOAP section/subsection to guide summary content and style. **BART+ADAPT** integrated adapter modules - small bottleneck feed-forward networks - into each transformer layer, adopted from (Bapna and Firat, 2019). To enable SOAP note generation, full finetuning was applied. **BART+CA** adds a second cross-attention (CA) layer after the standard CA layer in each decoder block. While the default CA layer shares parameters across all sections, this additional layer uses separate parameters for attending to SOAP information. This enables the model to re-attend over the encoder output with a section-specific focus. These parameters are randomly initialized and learned during finetuning. Among the BART architectures, the BART+CA combination outperformed other modifications by a range of just 0.4–5.5 points in 3 of 4 applied metrics (ROUGE-L, UMLS(F-1), Isin). Compared to the baseline BART and GPT 3.5, the BART+CA modification outperformed other models in 2 of 4 metrics (UMLS(F-1), Direc(F-1)) by a range of 0.7 - 15.6 points (Ramprasad et al., 2023). These findings imply that a specialization on an architectural level for a task can benefit a model's performance.

3 Methodology

Figure 1 shows the methodological process steps.

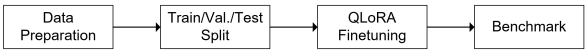


Figure 1: Methodology process steps

3.1 Dataset and Data Preparation

This study employed the MedSynth dataset, created by Mianroodi et al. (2025). The authors leveraged data from multiple sources, such as filed insurance claims from the IQVIA PharMetrics Plus dataset, and existing dialogue datasets, such as Aci-Bench (Yim et al., 2023), NoteChat (Wang et al., 2024), and PriMock57 (Papadopoulos Korfiatis et al., 2022). Despite the highlighted categorical skewness of the distribution of ICD-10 codes, the authors have selected a uniform sampling method, resulting in ca. 2,000 selected ICD-10 codes for English dialogue-note pair generation. The dialogues were synthetically generated with multiple AI agents, employing GPT-4o with Chain-of-Thought and In-Context Learning. The incident descriptions from the selected insurance claims were taken as an input for the GPT4o model, in accordance with the instruction for generating a dialogue. Through this, ca. 10,000 dialogue-note pairs were created, where each ICD-10 code is generally represented 5 times.

The dataset version retrieved in Feb. 2026 comprises 10,240 dialogue-note pairs in total, featuring 2,037 ICD-10 code classes. After cleaning and preparation, the final dataset comprised of 10,030 records in 2,006 ICD-10 code classes. For the later training of multi- and swarm-agentic AI systems, the SOAP notes have been split via regex in their 4 respective dimensions. The evaluation dataset was constructed by randomly (all random selections employ a fixed seed) selecting one sample per ICD-10 code, resulting in 2,006 test instances (20%). This ensures uniform code-level coverage in the final evaluation. The remaining samples are randomly shuffled split into training and validation sets. Specifically, 802 samples (8%) are used for validation to support model selection, while the remaining 7,222 samples (72%) are used for training. This splitting strategy balances the need for representative evaluation across clinical categories with maintaining sufficient training data for reliable finetuning.

The records from the test set, serving as ground-truth SOAP notes, were scored directly against their source dialogues to establish a lexical-overlap baseline. All notes were run directly against their source dialogues without model inference, quantifying the inherent textual distance between raw dialogue and structured SOAP output. This baseline is intentionally conservative: it does not represent a competitive note-generation system, but rather provides a lower-bound reference that quantifies how much information must be reorganized, paraphrased, and inferred to transform raw dialogue into a structured SOAP note. Direct numerical comparison with dedicated baselines from prior work such as Ramprasad et al. (2023) or Li et al. (2025) is informally enabled through the shared use of ROUGE and BERTScore, although differences in datasets, target schemas, and reference quality preclude strict equivalence.

3.2 Benchmark Metrics

The following metrics were applied:

- **ROUGE**: This metric was proposed by Lin (2004) and stands for *Recall-Oriented Understudy for Gisting Evaluation*. This metric evaluates the overlap between generated and reference texts using n-gram matching. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap respectively, while ROUGE-L captures sentence-level structural similarity without requiring consecutive n-gram matches. The metric reports precision, recall, and F1, all of which were reported.

- **BERTScore**: BERTScore was proposed by Zhang et al. (2020). BERTScore computes cosine similarity between contextual embeddings of tokens in the generated and reference texts, enabling it to capture semantic equivalence even when different words are used. This study leverages the RoBERTa-large model (Liu et al., 2019) to compute the score. Furthermore, the metric reports precision, recall, and F1, of which all were reported.

The combination of ROUGE and BERTScore provides complementary evaluation perspectives: ROUGE captures lexical overlap through surface-level n-gram matching, while BERTScore assesses semantic similarity through contextual embeddings, recognizing paraphrases that ROUGE would miss. Both metrics are widely adopted in clinical note generation research. This enables a direct comparison with prior work such as Li et al. (2025); Mianroodi et al. (2025) and others.

Recent shared tasks on clinical text generation, such as RRG24 in Xu et al. (2024), have employed broader metric suites including F1-RadGraph and F1-CheXbert for radiology report generation. These metrics are specifically tailored to radiology and rely on radiology-domain labelers, whereas the *Discharge Me!* subtask instead used metrics such as MEDCON and AlignScore for discharge summary generation. Neither suite transfers directly to general SOAP note evaluation across the full ICD-10 spectrum, where no analogous domain-specific labelers are established. ROUGE and BERTScore remain the metrics most consistently reported in the SOAP-generation literature and enable direct comparison with prior work (Li et al., 2025; Ramprasad et al., 2023).

3.3 LLM Selection and Finetuning

3.3.1 LLM selection

All architectures in this study employed the Ministral 3 model series (Liu et al., 2026). Specifically, the 3B and 8B reasoning model versions *mistralai/Ministral-3-3B-Reasoning-2512* and *mistralai/Ministral-3-8B-Reasoning-2512* were used for finetuning and benchmarking in this study. The model family was published by Mistral AI in Jan. 2026. Its strong performance on the MMLU-Redux benchmark (Gema et al., 2025), a benchmark for multitask language understanding where LLMs are confronted with questions from all kinds of domains, made the model an ideal choice. Additional consideration was given to its native support for structured output generation and level 3 agentic AI capabilities. The model employs a decoder-only transformer architecture and supports context lengths of up to 256k tokens; sufficient for processing medical dialogues with generative instructions.

3.3.2 Quantization

To still leverage the model's capabilities while improving memory efficiency, the quantization technique *Quantized Low Rank Adaptation* (QLoRA) was used. Introduced by Dettmers et al. (2023), QLoRA combines *Normal Float 4* (NF4) quantization with *Low Rank Adaptation* (LoRA) (Hu et al., 2022), enabling parameter-efficient finetuning by keeping base weights frozen and training small low-rank adapter matrices instead. It further reduces memory usage through double quantization.

Quantization represents model weights in lower-precision formats, trading off compression and accuracy. While Ministral 3 natively uses *bfloat16*, alternatives such as INT8, INT4, FP8, and NF4 offer varying memory savings. Given the focus on maintaining quality under strong compression, NF4 was selected.

QLoRA exploits the near-normal distribution of pretrained weights by aligning quantization levels with weight density, minimizing information loss at 4-bit precision. Double quantization further compresses the per-block scaling factors, reducing overhead to $\sim$0.127 bits per parameter. During training, only LoRA adapters are updated while 4-bit weights are temporarily dequantized to bfloat16; paged optimizers stabilize memory usage on single-GPU setups.

This study applied LoRA hyperparameters as listed in Table 1:

Parameter	Value	Description
Rank (r)	64	Dimension of the low-rank matrices
Alpha (α)	32	Scaling factor for LoRA updates
Dropout	0.05	Dropout probability
Target modules	q/k/v/o/gate/ up/down_proj	Transformer layers adapted
Bias	None	No bias params trained
Task type	CAUSAL_LM	Language modeling objective

Table 1: LoRA hyperparameter configuration

Preliminary runs with r=16 and r=32 showed marginally lower validation performance; r=64 was selected as the best tradeoff.

3.3.3 Finetuning

The finetuning of the respective base models with 3B and 8B parameters was conducted on the training split with 7,222 observations, and tested against the validation split of 802 observations for each epoch. Table 2 shows the applied hyperparameters for training for all architectures.

The finetuning results suggest that the models reach their global optimum within 3 epochs. Figure 2 displays the performance of the training and validation for the single-LLM architecture, which

Parameter	Value
Learning rate	2×10^{-4}
Epochs	3
Per-device batch size	2
Gradient accumulation steps	4
Effective batch size	8
Optimizer	AdamW
Learning rate scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Max sequence length	4096
Gradient checkpointing	Enabled
Precision	bf16

Table 2: Finetuning hyperparameter configuration

can be considered representative for other models, employed in different architectures, as other finetuning runs show similar results. Training was conducted on 3× NVIDIA A100-40GB GPUs.

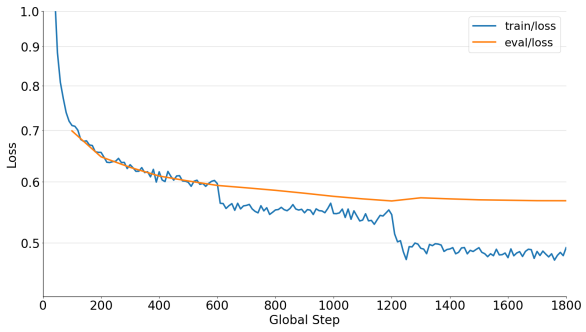


Figure 2: Single-LLM setup 8B model training and validation performance

3.3.4 Preservation of Base Model Capabilities

A potential concern when finetuning models for downstream agentic use is the loss of general instruction-following capability that pre-trained models acquire during instruction tuning. The QLoRA setup adopted here was specifically chosen to mitigate this risk. Three properties of the configuration are relevant: (1) All base model weights remain frozen in 4-bit NF4 quantization throughout training; gradient updates are applied only to the rank-64 LoRA adapter matrices, which account for ca. 2.9% (3B) and ca. 2.3% (8B) of total parameters — a range consistent with standard QLoRA configurations applied to comparable base-model scales (Dettmers et al., 2023; Li et al., 2025) (2) Each adapter parameterizes its weight update as a low-rank decomposition ($\Delta W = BA$, with rank r = 64), restricting the modification of every targeted projection to a 64-dimensional subspace; this structural constraint bounds how far the adapted layer can deviate from the frozen pre-trained projection. (3) During inference, agents continue to receive role-specific system prompts (defining role, expected output structure, and clinical scope) — finetuning thus complements, rather than replaces, prompt-based agent specialization. Empirical support for this design choice can be found in prior clinical NLP work that combines QLoRA adapters with structured prompting for SOAP-related tasks (Li et al., 2025). An alternative design — for example, orchestrating off-the-shelf instruction-tuned LLMs through prompts alone — represents a valid and complementary configuration, and is positioned as a future-work direction in Section 6.

3.4 Agentic AI Architectures

Before the architectures are described, it is worth noting that for both the multi-agentic and swarm-agentic AI architectures, trained LoRA adapters were applied via hot-swap to a single base model during inference, significantly reducing VRAM demands as only one base model needs to be held in memory. For all architectures, all LLMs were either equipped with 3B or 8B parameters.

Note that this study employs role-specialized adapters within a fixed pipeline rather than prompt-orchestrated agents over a single instruction-tuned base model. Coordination between agents is structural — information flow is hard-wired through the architecture (Figures 4, 5) — not negotiated at inference. This design isolates architectural decomposition as the experimental variable while holding base model and training recipe constant; alternative prompt-driven setups represent a complementary design space (see Section 6).

3.4.1 Single-LLM Architecture

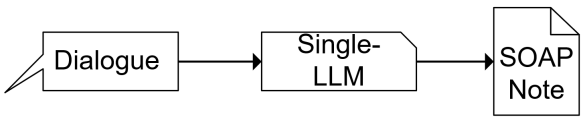


Figure 3: Single-LLM architecture visualized

The single-LLM setup featured a finetuned model, trained to compose an entire SOAP note. During training, the entire SOAP note for the individual case was visible. This enabled the model to capture the holistic format and structure of the SOAP-schema. For inference, the trained adapter

was applied. The purpose of the single-LLM setup is to serve as a controlled comparison against agentic AI architectures with multiple AI agents. Figure 3 visualizes the single-LLM architecture.

3.4.2 Multi-Agentic AI Architecture

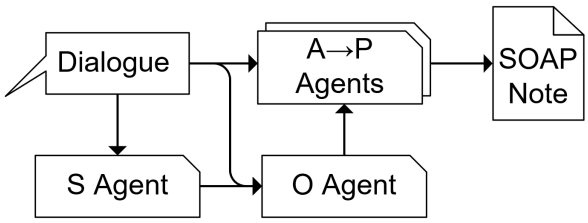


Figure 4: Multi-agentic AI architecture visualized

The multi-agentic AI setup featured 4 agents in sequential order, each one finetuned in their respective domain. Depending on their relative position in the order of the SOAP schema, the models receive the dialogue, as well as the output of the previous AI agents. For the final output, all dimension-outputs were concatenated in the appropriate SOAP order. Figure 4 visualizes the multi-agentic AI architecture.

For training, the SOAP notes have been separated in their respective domains, to control the access of information for each SOAP stage. Every domain-respective agent received a specialized system prompt (e.g., defining role, expected output format, and relevant clinical concepts) to ensure dimension-specific information sensitivity for phrases and concepts described in the dialogue.

3.4.3 Swarm-Agentic AI Architecture

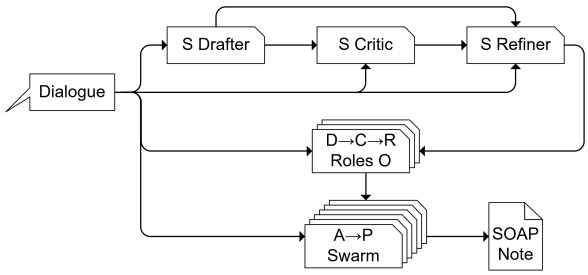


Figure 5: Swarm-agentic AI architecture visualized

The swarm-agentic AI setup featured 12 agents in sequential order, with 3 agents per SOAP dimension acting as drafter, critic, and refiner. For every dialogue input, the *drafter* composes a SOAP note for the respective dimension. The *critic* then judges the drafter's output against the original dialogue, based on structure and criteria learned through finetuning. The *refiner* refines the draft based on the dialogue, drafter output, and critic output, and outputs the final SOAP-dimension note. All dimension outputs were concatenated in the appropriate SOAP order. Figure 5 visualizes the swarm-agentic AI architecture.

During training, each of the 12 agents received a specialized system prompt (e.g., defining role, expected output format, and relevant clinical concepts) to ensure dimension specialization through role understanding and sensitivity to relevant phrases and concepts. The drafter was trained on dialogue-dimension pairs with the ground-truth SOAP dimension as gold text, analogous to the multi-agent training. The critic relied on synthetic data: a cross-paired imperfect draft is created by swapping the SOAP dimension with one from a different consultation sharing the same ICD-10 code, and a programmatic function generates a synthetic critique comparing that draft against the ground-truth SOAP note and current dialogue. The critic thus learns to generate a critique from dialogue and imperfect draft. The refiner is then trained on the same cross-pairs, receiving dialogue, imperfect draft, and synthetic critique as input, with the gold-standard SOAP dimension as target. For inference, the 12 trained adapters were applied via hot-swap to the same base model.

4 Results

4.1 Quantitative Results

For benchmarking, the test set with 2,006 observations, representing one ICD-10 code class each, was employed. The computed quality metrics allowed a per-case review for every setup and every agent. Table 3 features the combined performance metrics over all architecture respective agents, where the top performing architectural metric is formatted in **bold**, while the 2nd best performing architecture metrics are underlined.

The single-LLM architecture achieves the strongest performance, ranking first across all metrics, with the 8B variant marginally outperforming the 3B variant. Both share nearly identical BERTScores ($\Delta = 0.002$). The multi-agent and swarm-agent setups follow closely, with Multi 8B consistently outperforming Swarm 8B across all metrics, while Single 3B holds the second-ranked position in all four metrics. The performance differences across all finetuned architectures are marginal, with the largest gap being 0.042 in ROUGE-L between Single 8B (0.628) and Multi

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Baseline	0.461	0.227	0.326	0.825
Single 3B	0.771$_{\pm.002}$	0.543$_{\pm.003}$	0.615$_{\pm.003}$	0.928$_{\pm.001}$
Single 8B	**0.778**$_{\pm.002}$	**0.554**$_{\pm.003}$	**0.628**$_{\pm.003}$	**0.930**$_{\pm.001}$
Multi 3B	0.753$_{\pm.003}$	0.519$_{\pm.003}$	0.586$_{\pm.004}$	0.921$_{\pm.001}$
Multi 8B	0.760$_{\pm.002}$	0.530$_{\pm.003}$	0.598$_{\pm.003}$	0.922$_{\pm.001}$
Swarm 3B	0.752$_{\pm.003}$	0.522$_{\pm.004}$	0.591$_{\pm.004}$	0.920$_{\pm.001}$
Swarm 8B	0.754$_{\pm.003}$	0.526$_{\pm.003}$	0.594$_{\pm.004}$	0.921$_{\pm.001}$

Table 3: Test set F1 results. Baseline denotes lexical-overlap on ground-truth SOAP notes. ±values denote 95% bootstrap CI half-widths (10,000 resamples).

3B (0.586).

To evaluate statistical significance, paired t-tests and Wilcoxon signed-rank tests were conducted on per-case scores across all 2,006 test instances, with Cohen's d as effect size. The single-LLM architecture significantly outperforms both the multi-agent and swarm-agent setups across all metrics (all $p < 0.001$), with medium BERTScore effect sizes (d = 0.67–0.80 for 8B; d = 0.65–0.70 for 3B). The multi-agent vs. swarm-agent gap is negligible (BERTScore Δ = 0.0007, d = 0.050; Wilcoxon p = 0.161). For model scale, 8B variants show small but significant ROUGE improvements over 3B ($d \approx 0.15$–0.28), while BERTScore differences exhibit negligible effect sizes for both multi- and swarm-agent settings ($d \leq 0.060$; p = 0.149 and p = 0.007 respectively), suggesting that architectural choice dominates over parameter count in practical terms.

4.2 Qualitative Analysis

A stratified inspection of 100 generated notes, drawn uniformly across ICD-10 categories, revealed three recurring patterns across all finetuned configurations.

Demographic hallucinations in the Subjective dimension. All architectures frequently produced patient demographic attributes (age, gender, ethnicity) absent from the source dialogue. These attributes appear consistently in MedSynth's reference notes without corresponding dialogue evidence, and models reproduced them as a learned prior rather than through factual extraction — a dataset-level bias that architectural decomposition did not mitigate.

Empty Objective dimensions. Approximately 15% of reference notes contained no Objective content, reflecting MedSynth's simulated telemedicine scenarios. While all architectures correctly reproduced empty sections in these cases, multi- and swarm-agent variants occasionally hallucinated brief filler text ("examination not performed", "vitals not recorded"), marginally reducing reference alignment.

Error propagation in sequential pipelines. In the swarm setup, errors introduced at the drafter stage — such as fabricated medication names — tended to persist through critic and refiner stages. The critic, trained on synthetic cross-paired critiques, produced plausible critique text but failed to catch factual issues requiring external clinical knowledge. This aligns with the negligible BERTScore gap between multi- and swarm-agent variants (Section 4.1) and explains why additional refinement stages did not yield gains.

Together, these observations suggest that performance differences across architectures are driven less by architectural reasoning capacity than by shared dataset priors and drafter error propagation — limitations unlikely to be resolved by further architectural complexity.

5 Discussion

The results show that overall the finetuned models substantially outperform the lexical-overlap baseline, which measures the inherent textual overlap between raw dialogues and ground-truth SOAP notes without any model inference. This confirms the value of task-specific finetuning.

Figure 6 shows the F1 BERTScore delta between architectures; among finetuned models, the absolute maximum is 0.008 and the minimum is 0 — indicating that 3B models perform competitively with larger and more complex configurations alike. A plausible explanation is that SOAP note generation, as a structured reorganization and paraphrase task, does not expose the complexity that multi-agent decomposition is designed to exploit: a single finetuned model captures the full dialogue-to-

Δ	Base	Single 3B	Single 8B	Multi 3B	Multi 8B	Swarm 3B	Swarm 8B
Base		-0.103	-0.105	-0.096	-0.097	-0.096	-0.096
Single 3B	0.103		-0.002	0.007	0.006	0.008	0.007
Single 8B	0.105	0.002		0.009	0.008	0.009	0.009
Multi 3B	0.096	-0.007	-0.009		-0.001	0.001	0.000
Multi 8B	0.097	-0.006	-0.008	0.001		0.002	0.002
Swarm 3B	0.096	-0.008	-0.009	-0.001	-0.002		0.000
Swarm 8B	0.096	-0.007	-0.009	0.000	-0.002	0.000	

Figure 6: Delta for F1 BERTScore benchmark results comparison (row - column)

note mapping end-to-end, without the coordination overhead and error propagation inherent to pipeline architectures. We therefore interpret the flat architectural performance not as a failure mode but as an empirical answer — agentic decomposition is not warranted for this class of tasks. This counterpoints the broader trend of defaulting to multi-agent orchestration in clinical NLP and motivates the targeted question outlined in Section 6: identifying the specific task properties — multi-specialty, longitudinal, multi-modal — under which decomposition can be expected to pay off.

6 Conclusion

This study compared three agentic AI architectures — single-LLM, multi-agentic, and swarm-agentic — for automated SOAP note generation from doctor–patient dialogues, using finetuned Ministral 3 models (3B and 8B) on the MedSynth dataset. All finetuned architectures substantially outperformed the lexical-overlap baseline, underscoring the value of task-specific finetuning in clinical documentation. Performance differences among the three approaches were minimal, however, with the single-LLM setup leading in all 4 metrics. These findings suggest that increasing architectural complexity through multi-agent or swarm decomposition does not provide meaningful gains over a single finetuned model for this task. Notably, the 3B and 8B variants performed nearly identically, indicating that smaller models can handle structured clinical NLP tasks without compromising quality.

These findings carry practical implications: organizations seeking to automate clinical documentation can achieve competitive results with a single, compact finetuned model, reducing both infrastructure complexity and computational cost compared to multi-agent deployments.

Several limitations should be acknowledged. The evaluation relied exclusively on automated metrics (ROUGE, BERTScore), which may not fully capture clinical accuracy or factual correctness — dimensions that require expert human evaluation. Furthermore, the study was conducted on a single synthetic dataset, and generalizability to real-world clinical dialogues with noise from speech recognition remains to be validated.

As detailed in Section 4.2, the inference outputs reveal dataset-induced behavior — demographic hallucinations in the Subjective dimension and reproduction of empty Objective sections from MedSynth's telemedicine scenarios — that limits its practical utility independently of architectural choice, since key content decisions appear driven by pattern reproduction, not factual grounding.

Future work should (1) extend evaluation to authentic clinical transcripts with acoustic and disfluency noise; (2) incorporate human expert assessment of factual correctness, completeness, and clinical utility; (3) test additional base-model families and instruction-only orchestrations to disentangle architectural from base-model effects; (4) explore broader, factuality-oriented metrics (e.g., RRG24-style entity-level scores adapted to SOAP content (Xu et al., 2024)); and (5) investigate whether more complex clinical tasks — multi-specialty consultations, longitudinal patient records, or multi-modal inputs combining dialogue with imaging or lab data — provide the conditions under which agentic decomposition begins to pay off.

7 Acknowledgements

This paper is dedicated to my loving partner Anna-Isabel Steffens for her unwavering belief in me.

My sincere gratitude to my project partner, Julia Jellinek, for her outstanding support in data engineering.

References

American Medical Informatics Association. 2024. TrendBurden-results-april-2024.

Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Sigall K. Bell, Tom Delbanco, Joann G. Elmore, Patricia S. Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G. Leveille, Thomas H. Payne, Rebecca A. Stametz, Jan Walker, and Catherine M. DesRoches. 2020. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Network Open*, 3(6):e205867.

Michael James Bommarito, Jillian Bommarito, and Daniel Martin Katz. 2025. What is an agent? a conceptual primer and history of agents and agentic AI.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: Efficient Finetuning of Quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, and 1 others. 2025. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ignetica Ltd. 2022. Assessing the burden of clinical documentation.

Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. 2025. Improving clinical note generation from complex doctor-patient conversation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 209–221. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ahmad Rezaie Mianroodi, Amirali Rezaie, Niko Grisel Todorov, Cyril Rakovski, and Frank Rudzicz. 2025. Medsynth: Realistic, synthetic medical dialogue-note pairs. *Preprint*, arXiv:2508.01401.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2023. SOAP notes. In *StatPearls*. StatPearls Publishing.

Sanjana Ramprasad, Elisa Ferracane, and Sai P Selvaraj. 2023. Generating more faithful and consistent SOAP notes using attribute-specific parameters. *Proceedings of the 8th Machine Learning for Healthcare Conference*, 8(219):631–649.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6594–6604.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lawrence L. Weed. 1964. Medical records, patient care, and medical education. *Irish Journal of Medical Science*, 39(6):271–282.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, and 1 others. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

VERICITE: Evaluating Sentence-Level Citation Faithfulness in Retrieval-Augmented Medical Question Answering

Yixian Ma **Bohao Chu** **Norbert Fuhr**
University Duisburg-Essen
{yixian@stud, bohao.chu, norbert.fuhr}@uni-due.de

Abstract

Retrieval-augmented generation (RAG) reduces hallucination in large language models by grounding outputs in retrieved evidence, but it does not guarantee that the resulting citations actually support the claims they accompany. We present VERICITE, a framework for evaluating citation faithfulness in retrieval-augmented medical QA. Our system retrieves PubMed abstracts via the NCBI E-utilities API, prompts LLMs to generate answers with inline citations, and then verifies each citation at the sentence level using a DeBERTa-v3-large NLI model. We evaluate four LLMs on 500 BioASQ questions at retrieval depths $k \in \{3, 5\}$, with extended experiments up to k=15 and an oracle setting with gold-standard documents. Only 27–41% of citation pairs are supported at the sentence level at $k \in \{3, 5\}$, with rates declining further at larger k. Under the oracle condition, answer quality improves but citation faithfulness does not substantially improve, suggesting that generation-side citation behavior contributes substantially to unfaithful citations.

1 Introduction

RAG (Lewis et al., 2020) has become standard practice in medical question answering (QA) because it grounds model outputs in external evidence (Xiong et al., 2024). However, the presence of a citation does not mean the cited source actually supports the claim. A model can attach a real PubMed article to a plausible-sounding statement even when the article says something only tangentially related—what Wallat et al. (2025) call "post-rationalized" citations. In medicine this is especially dangerous: a clinician may trust a cited claim precisely because it appears backed by evidence (Sackett et al., 1996).

We ask a simple question: when an LLM cites a retrieved PubMed abstract, does that abstract actually contain a sentence supporting the claim? VERICITE retrieves abstracts via PubMed E-utilities, prompts an LLM to generate a cited answer, and then checks each citation at the sentence level using NLI. We retrieve at the abstract level to preserve clinical context but verify at the sentence level for precise grounding. Our main contributions are:

1. A biomedical RAG pipeline with sentence-level NLI faithfulness evaluation. While ALCE (Gao et al., 2023) introduced sentence-level citation evaluation in open-domain QA, we adapt this paradigm to biomedical retrieval over PubMed abstracts.

2. An oracle retrieval experiment that approximates an upper-bound retrieval condition and analyzes how citation faithfulness changes under near-perfect recall. Gold-standard documents improve answer quality but not citation faithfulness, providing suggestive evidence that generation-side factors may contribute substantially to unfaithful citations.

3. A comparison of four LLMs across four retrieval depths, showing a precision–coverage trade-off where increasing k improves recall but degrades faithfulness.

2 Related Work

RAG and citation in medical QA. The ALCE benchmark (Gao et al., 2023) introduced sentence-level citation evaluation via citation recall and precision, finding that even strong models fail to support their citations about half the time; however, ALCE targets open-domain QA. MedCite (Du et al., 2025) compared citation strategies for biomedical QA, and the TREC BioGen shared task (Gupta et al., 2024, 2026b) has established benchmarks for evaluating biomedical generative retrieval with expert assessments. Most of this work focuses on producing citations; our focus is on verifying them at fine granularity in the biomed-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 753–759
July 3-4, 2026 ©2026 Association for Computational Linguistics

ical domain and on analyzing citation behavior under different retrieval conditions.

Citation faithfulness. Wallat et al. (2025) drew an important distinction between citation *correctness* (the source is real) and citation *faithfulness* (the source entails the claim), and found that most RAG citations are post-rationalized. Vladika et al. (2025) showed that even GPT-4o with retrieval augmentation lacks citation support for nearly half of its medical responses. BioACE (Gupta et al., 2026a) proposed automated biomedical citation evaluation and highlighted the difficulty of NLI-based assessment in this domain. Self-RAG (Asai et al., 2024) trains a single LM to critique its own retrieval via reflection tokens. We are closest to ALCE and Source-Checkup. ALCE evaluates citation quality in open-domain QA, while we adapt sentence-level evaluation to PubMed-based biomedical QA with both standard and oracle retrieval. Source-Checkup studies citation support in medical QA but lacks this retrieval comparison.

NLI for factual verification. NLI has been widely used for factual consistency checking (Maynez et al., 2020). We use DeBERTa-v3-large (He et al., 2023) fine-tuned on MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), ANLI (Nie et al., 2020), LingNLI, and WANLI.[1] This model was not trained on biomedical NLI data such as MedNLI (Romanov and Shivade, 2018); we discuss the implications in Section 6.

3 Method

Figure 1 shows the VERICITE pipeline.

3.1 Retrieval and Cited Answer Generation

Given a medical question, we retrieve the top-k PubMed abstracts using the NCBI E-utilities API (Sayers et al., 2022). The question is preprocessed by removing the leading interrogative word, preserving medical terminology for PubMed's Automatic Term Mapping. We retrieve a candidate pool of 200 results with title/abstract field restriction and keep the first k abstracts with complete metadata (Appendix A). Retrieval operates at the abstract level because isolated biomedical sentences often lack key context such as population, intervention, or outcome qualifiers. The abstracts are then passed to an LLM that generates a 4–6 sentence answer with inline document-level citations (Appendix B and E).

We intentionally use simple keyword-based retrieval (E-utilities) rather than dense retrieval, since our goal is to evaluate citation faithfulness across a range of recall levels—from low recall (standard retrieval) to near-perfect recall (oracle)—rather than to optimize retrieval itself.

Oracle retrieval. To separate retrieval from generation effects, we include an oracle condition that uses BioASQ's gold-standard PMIDs directly, bypassing search. This yields recall ≈ 1.0 (99.9%; a few gold PMIDs lack retrievable abstracts). By comparing standard and oracle conditions on the same models, we can examine whether substantially improved retrieval recall is associated with changes in citation faithfulness.

3.2 Sentence-Level Citation Faithfulness

Although citations are produced at the document level, we evaluate faithfulness at the sentence level. For each generated sentence that cites a document, we segment the cited abstract into individual source sentences and run NLI on every (generated claim, source sentence) pair. This is a sentence-to-sentence comparison, not a comparison between the generated sentence and the full abstract. Each source sentence is evaluated independently against the generated claim using the DeBERTa-v3-large NLI model described above.

We aggregate the per-sentence NLI verdicts into a single citation-level verdict for each (generated sentence, cited document) pair: **Supported** if at least one source sentence entails the claim; **Contradicted** if none entails but at least one contradicts; **Irrelevant** otherwise (all source sentences are neutral).

4 Experimental Setup

Data. We use 500 questions (excluding yes/no) from the BioASQ Task 13b training set (Tsatsaronis et al., 2015). BioASQ provides expert-authored ideal answers and gold-relevant PMIDs. Of these, 421 yield at least one abstract under standard retrieval—factoid (161), list (114), summary (146)—and are used for those conditions; per-type results are in Appendix D. The oracle condition uses all 500 questions.

[1] Checkpoint: `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` on Hugging Face.

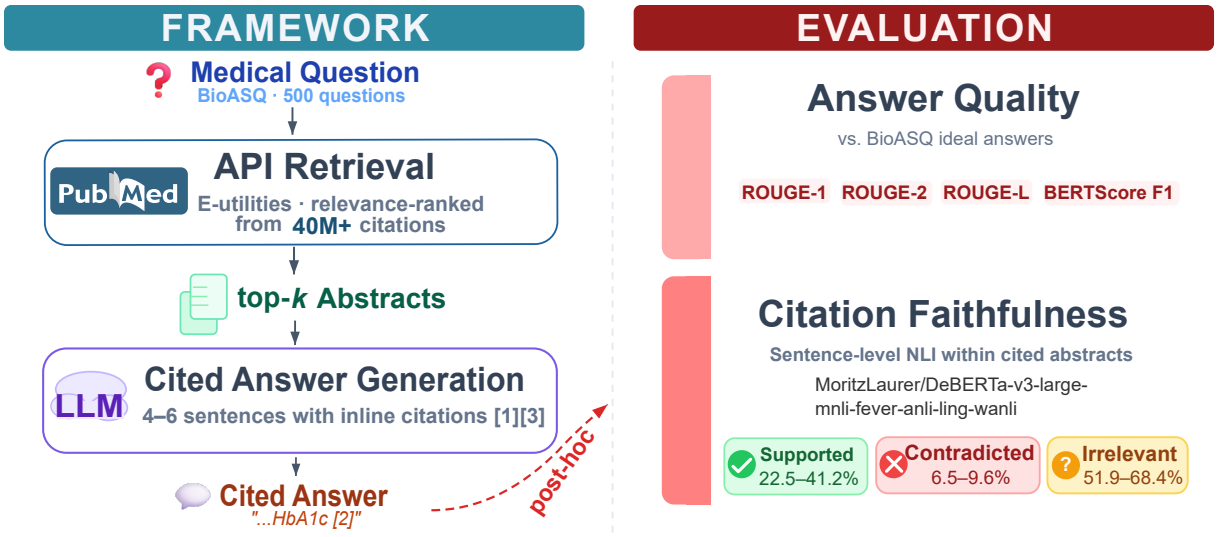


Figure 1: The VERICITE pipeline. **Left**: retrieval of top-k PubMed abstracts and cited answer generation. **Right**: post-hoc evaluation of retrieval quality, answer quality, and sentence-level citation faithfulness via NLI.

Models. We evaluate four LLMs, all given identical retrieval results and the same prompt: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-70B-Instruct (Dubey et al., 2024), GPT-4o-mini (OpenAI, 2024), and Claude-3.5-Haiku (Anthropic, 2024), accessed via API (Mistral through DeepInfra; others through OpenRouter). We include both a small open-weight model (Mistral-7B) and larger models to examine whether model scale affects citation behavior. All four run at $k \in \{3, 5\}$; Mistral-7B and Llama-3.1-70B also run at $k \in \{10, 15\}$. Temperature is 0.4 with nucleus sampling p=0.9.

Metrics. We report three groups: (1) *Retrieval*: Recall@k against BioASQ gold PMIDs. (2) *Answer quality*: ROUGE-1/2/L (Lin, 2004) and BERTScore F1 (Zhang et al., 2020) (`roberta-large`) against ideal answers. (3) *Citation faithfulness*: the fraction of (generated sentence, cited document) pairs classified as supported, contradicted, or irrelevant. Note that the ROUGE/BERTScore comparison is not entirely fair, since BioASQ ideal answers were written with access to a broader set of gold documents than our models see.

5 Results

Table 1 shows results for all models at $k \in \{3, 5\}$.

Retrieval. Recall@3 is 0.127 and Recall@5 is 0.162 over the 421 retained questions (79 returned no abstracts and were excluded). This low recall reflects two factors: BioASQ gold references are curated through expert literature review and may not rank highly in keyword search, and our E-utilities retrieval is deliberately simple. We treat this as a lightweight retrieval baseline rather than a competitive biomedical retriever; the oracle condition provides a high-recall counterpart for analyzing citation behavior under substantially improved retrieval coverage.

Citation faithfulness. Under our sentence-level NLI criterion—which may underestimate true faithfulness when models paraphrase or merge information across source sentences (Section 6)—only 29.7–41.2% of citation pairs at k=3 are supported. The majority (51.6–61.7%) are irrelevant: the cited abstract relates to the question topic but no individual source sentence entails the specific claim. Contradictions account for 6.6–9.5%, a smaller but non-trivial fraction in a medical context. These patterns are consistent with Wallat et al. (2025), who found most RAG citations to be post-rationalized.

Among models, Llama-3.1-70B achieves the highest support rate (41.2% at k=3), followed by Claude-3.5-Haiku (38.9%), GPT-4o-mini (37.8%), and Mistral-7B (29.7%). The gap between Mistral-7B and the other models suggests that model capacity may play a role, though we cannot isolate scale from other factors (training data, instruction tuning). Across both retrieval depths, GPT-4o-mini shows the lowest contradiction rate (6.5% at k=5), suggesting relatively conservative citation behavior.

Model	k	Retr. Rec.	Citation Faithfulness Sup.↑	Con.↓	Irr.↓	Answer Quality R-1	R-2	R-L	BS
Mistral-7B	3	.127	.297	.086	.617	.251	.091	.171	.862
	5	.162	.274	.078	.648	.258	.097	.177	.864
Llama-3.1-70B	3	.127	**.412**	.069	.519	.259	.107	.185	.863
	5	.162	.391	.066	.543	**.269**	**.118**	**.196**	**.865**
GPT-4o-mini	3	.127	.378	.066	.556	.249	.090	.171	.862
	5	.162	.360	**.065**	.575	.256	.097	.175	.865
Claude-3.5-Haiku	3	.127	.389	.095	**.516**	.249	.088	.171	.860
	5	.162	.364	.080	.556	.257	.096	.178	.863

Table 1: Main results on BioASQ (421 questions). ↑/↓: higher/lower is better. Best value per column in **bold**. BS: BERTScore F1 (`roberta-large`).

Model	k	Rec.	Sup.	Irr.	R-L	BS
Mistral	3	.127	.297	.617	.171	.862
	5	.162	.274	.648	.177	.864
	10	.207	.240	.667	.178	.865
	15	.226	.227	.684	.181	.865
	orac.	*.999*	*.225*	*.678*	*.212*	*.876*
Llama-70B	3	.127	.412	.519	.185	.863
	5	.162	.391	.543	.196	.865
	10	.207	.364	.560	.213	.869
	15	.226	.334	.590	.223	.871
	orac.	*.999*	*.381*	*.533*	*.256*	*.882*

Table 2: Extended depth ($k \in \{3, 5, 10, 15\}$; 421 questions) and oracle with gold PMIDs (500 questions). Support rates decrease with k; oracle retrieval does not improve faithfulness.

Extended retrieval depth. Table 2 extends the analysis to $k \in \{3, 5, 10, 15\}$ for Mistral-7B and Llama-3.1-70B. As k increases from 3 to 15, Recall@k rises to 0.226 and answer quality improves modestly, but support rates drop monotonically: from 29.7% to 22.7% for Mistral, and from 41.2% to 33.4% for Llama. The reason is straightforward: models cite more documents per answer at larger k (2.5 to 5.5 unique docs for Mistral), but many of the additional citations point to documents only loosely related to the specific claim. This constitutes a precision–coverage trade-off: increasing k improves retrieval recall but degrades citation precision.

Oracle retrieval. With recall of 99.9%, citation faithfulness does not substantially improve under oracle retrieval: support drops to 22.5% for Mistral-7B (vs. 29.7% at k=3) and remains comparable for Llama at 38.1% (vs. 41.2%). Meanwhile, answer quality improves clearly (ROUGE-L: .171 → .212 for Mistral; .185 → .256 for Llama). This divergence—better answers but not substantially better citation faithfulness—is one of our central empirical observations, suggesting that improved retrieval evidence alone may be insufficient to ensure faithful citation attribution.

The oracle setting differs from standard retrieval in both document count and question coverage, complicating strict causal interpretation. Increased document availability may encourage models to cite more sources, thereby reducing sentence-level support rates under our evaluation protocol. Nevertheless, the comparison with k=15 partially mitigates this concern, since Llama cites a comparable number of documents yet still achieves lower support rates than under oracle retrieval.

6 Conclusion

We presented VERICITE, a pipeline for evaluating citation faithfulness in retrieval-augmented medical QA. Across four LLMs at $k \in \{3, 5\}$, only 27–41% of citations are supported under sentence-level NLI evaluation; the majority are irrelevant. These rates generally decline at larger k and do not substantially improve under oracle retrieval. The oracle experiment reveals a notable divergence: providing models with expert-selected documents improves answer quality but does not substantially improve citation faithfulness. These findings suggest that retrieval quality alone may be insufficient to ensure faithful citation behavior, and that generation-side citation behavior likely plays an important role. For medical RAG systems, this means improving retrieval alone is unlikely to produce trustworthy citations; generation-side interventions such as citation-aware training or post-hoc verification are needed.

Limitations

First, we rely on a general-domain NLI model (DeBERTa-v3-large) not trained on biomedical text. Domain-specific models trained on MedNLI (Romanov and Shivade, 2018) might yield different re-

sults, and Gupta et al. (2026a) have highlighted the difficulty of NLI in biomedical settings. Without human validation, our faithfulness estimates should be treated as approximate; annotating a sample of NLI verdicts is our most important next step.

Second, our sentence-level entailment approach has known failure modes. When an LLM synthesizes information from multiple source sentences, no single source sentence may fully entail the combined claim, leading to false negatives. A generated sentence may also contain multiple sub-claims that we do not decompose. These factors mean our evaluation likely *underestimates* true faithfulness. A non-zero decoding temperature (0.4) may further increase paraphrasing diversity, reducing literal overlap with source sentences and lowering NLI entailment rates.

Third, the oracle experiment should be interpreted cautiously. In addition to near-perfect retrieval recall, the oracle setting also differs from the standard retrieval condition in both document count and question coverage, making strict causal interpretation difficult. As a result, the observed divergence between answer quality and citation faithfulness provides suggestive rather than definitive evidence that retrieval quality alone is insufficient to guarantee faithful citation behavior. A fully controlled comparison—for example, restricting oracle retrieval to the same question subset and matched document counts—would provide a stronger test.

Fourth, we evaluate on a single dataset (BioASQ) and retrieve only abstracts; full-text articles might provide sentence-level support that abstracts lack. Finally, API-based models may change across versions; we report exact model identifiers and fix decoding parameters to mitigate this.

Future Work

Our most pressing next step is human annotation of NLI verdicts to quantify evaluation reliability and compare the general-domain DeBERTa model against biomedical alternatives. We also plan to explore multi-sentence evidence aggregation (matching claims against groups of source sentences) to reduce false negatives, and to investigate dense retrieval and query expansion to improve recall without degrading citation precision.

Ethics Statement

This work evaluates citation reliability in medical QA and is not intended as a patient-facing system. AI-generated medical content should not substitute professional clinical judgment. All data are from publicly available sources.

References

Anthropic. 2024. Claude 3.5 model card.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Jiawei Du et al. 2025. MedCite: Towards faithful and traceable citation in biomedical question answering. *arXiv preprint arXiv:2501.07653*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2026a. BioACE: An automated framework for biomedical answer and citation evaluations. *arXiv preprint arXiv:2602.04982*.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of TREC 2024 biomedical generative retrieval (BioGen) track. In *Proceedings of the Thirty-Third Text REtrieval Conference (TREC 2024)*. ArXiv:2411.18069.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2026b. Overview of TREC 2025 biomedical generative retrieval (BioGen) track. *arXiv preprint arXiv:2603.21582*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*. ArXiv:2111.09543.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for

knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

OpenAI. 2024. GPT-4o system card.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72.

Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D206–D214.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Juraj Vladika et al. 2025. Source-checkup: A framework for evaluating source attribution in LLM-generated medical responses. *arXiv preprint arXiv:2501.09940*.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 22–32.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Retrieval Details

We submit the cleaned query to `esearch` with relevance ranking, restricting matches to title and abstract fields (`[tiab]`) and requiring a non-empty abstract (`hasabstract`). We retrieve 200 candidates via `efetch` and keep the first k with complete metadata. If the strict query returns fewer than k results, we relax the field restriction to allow Automatic Term Mapping, then fall back to the raw question.

For the oracle, we fetch abstracts using gold PMIDs directly. Recall is 99.9% rather than 100% because a few gold PMIDs lack abstracts. On average, models cite 5.9 (Mistral) and 5.3 (Llama) unique documents per answer under oracle.

B Generation Details

Each model receives the question and retrieved abstracts numbered $[1], \ldots, [k]$. The prompt includes a system instruction requiring the model to answer using only the provided documents, cite every sentence, and use multiple sources; two few-shot demonstrations; and the numbered abstracts and question (Appendix E). All models use the same prompt and documents. All four are accessed via API (Mistral-7B through DeepInfra; others through OpenRouter). Temperature is 0.4, p=0.9.

C Dataset Notes

Of 500 BioASQ questions, 421 yield at least one abstract under standard retrieval; the other 79 returned no usable abstracts. Recall@k ranges from 0.127 (k=3) to 0.226 (k=15). The oracle covers

Model	Type	Rec.	Sup.	Irr.	R-L	n
Mistral	Factoid	.108	.302	.603	.163	161
	List	.068	.274	.647	.146	114
	Summary	.195	.308	.610	.200	146
Llama	Factoid	.108	.421	.507	.179	161
	List	.068	.380	.551	.156	114
	Summary	.195	.425	.510	.214	146

Table 3: Citation faithfulness by question type at k=3. List questions show the lowest faithfulness, partly due to lower retrieval recall.

all 500 questions with recall of 0.999. We use questions from the Task 13b training set (2025 release), which accumulates 5,389 questions from previous years with expert-curated gold answers and relevant PMIDs.

D Results by Question Type

Table 3 breaks down citation faithfulness by question type at k=3. Summary questions show the highest support rates (30.8% for Mistral, 42.5% for Llama), likely because summary-type evidence is broader and easier to match. List questions show the lowest faithfulness (27.4% and 38.0%), partly due to low retrieval recall (6.8%). The ranking is consistent across both models.

E Prompt Template

```
You are a medical QA assistant. Below
↪  are
numbered source documents retrieved from
PubMed literature.

RULES:
1. Answer using ONLY the source
↪  documents.
2. Write between 4 and 6 sentences.
3. EVERY sentence MUST end with
↪  citation(s)
   in the format [1], [2][3], or [2,3].
4. Numbers refer to the document
↪  numbers.
5. Cite the most relevant supporting
↪  documents.
6. Do NOT add facts not in the sources.
7. If sources are insufficient, say so.

[Two few-shot demonstrations omitted]

Source documents:
[1] <abstract of document 1>
...
[k] <abstract of document k>

Question: <question>
Answer:
```

Overview of the Medical Decision Extraction, Analysis, and Classification Task (MedExACT) of BioNLP 2026

Mohamed Elgaar,[1] Jiali Cheng,[1] Nidhi Vakil,[1] Mehrnaz Sadrolashrafi,[2]
Mitra Mohtarami,[3] Adrian Wong,[2] Hadi Amiri,[1,4] Leo A. Celi[4]
[1]UMass Lowell, [2]BIDMC, [3]Anselm College, [4]MIT
{melgaar,jcheng2,nvakil,hadi}@cs.uml.edu, {msadrola,awong7}@bidmc.harvard.edu
mmohtarami@anselm.edu, lceli@mit.edu

Abstract

This paper presents an overview of the Medical Decision Extraction, Analysis, and Classification task (MedExACT) of BioNLP 2026. The focus of this task is the extraction and labeling of medical decisions in ICU discharge summaries. The task is built on MedDec, a MIMIC-III-based dataset of 451 expert-annotated summaries, and asks systems to extract and classify spans of text that contain medical decisions according to the decision categories defined in the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). The official ranking combines span F1 and token F1 with a worst-group robustness metric computed over sex, race, and English-proficiency subgroups. MedExACT attracted broad international interest, with 130 official submissions from 36 teams comprising about 60–100 participants, and has improved information extraction performance by nearly 15% over the previous state of the art. The submitted systems predominantly use long-context encoder models, ensemble decoding, boundary-refinement modules, and robustness-aware training or model selection, with the best submitted run reaching a final fairness-based F1 of 0.596.

1 Introduction

Clinical notes often document key medical decisions concerning diagnoses, medications, procedures, follow-up plans, and other decisions that shape downstream care. These decisions are central to clinical reasoning, but they are embedded in long free-text narratives that make them difficult to analyze at scale.

Prior studies have investigated joint span detection and classification for clinical outcomes (Abaho et al., 2021), weakly supervised extraction of medical evidence from clinical texts (Liu et al., 2021), and structured event extraction using sequence-to-sequence frameworks (Ma et al., 2023).

MedDec (Elgaar et al., 2024) introduced the first expert-annotated benchmark focused on extracting medical decisions from discharge summaries, grounding the task in the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016) and in the MIMIC-III critical care corpus (Johnson et al., 2016). The same data has also supported interactive downstream tooling for extraction, visualization, and annotation of medical decisions (Elgaar et al., 2025).

MedExACT 2026 extends the above efforts into a BioNLP shared task. Participants are asked to identify contiguous decision spans in ICU discharge summaries and assign each span one of the shared-task labels. Unlike other valuable information extraction resources such as CLIP (Mullenbach et al., 2021), which focus on physician action items, MedExACT targets a decision taxonomy that includes diagnostic assessments, therapeutic plans, and evaluation statements. The task averages standard performance (span- and token-level F1s) with a worst-group robustness metric across nine subgroups defined by sex, race, and English proficiency. This explicitly incentivizes fairness-aware system design instead of treating fairness as a post-hoc diagnostic. The task drew 130 submissions from 36 teams. The winning system scored 0.596. Top-performing approaches relied on encoder-based models with ensembles, boundary refinement, and category-specific decoding. LLMs performed well as augmentation or reranking components but poorly as standalone span extractors, consistent with prior MedDec findings.

This overview paper documents the task formulation and official evaluation approach, introduces the public baseline and toolkit distributed with the task, summarizes the systems submitted by participating teams and the modeling patterns that emerged from those submissions, and analyze the systems on exact span extraction, long-document processing, and subgroup-aware evaluation.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 760–769
July 3-4, 2026 ©2026 Association for Computational Linguistics

Decision Category	Description	Examples
Contact related	Decision regarding admittance or discharge from hospital, scheduling of control and referral to other parts of the healthcare system	Admit, discharge, follow-up, referral
Gathering information	Decision to obtain information from other sources than patient interview, physical examination and patient chart	Ordering test, consulting colleague, seeking external information
Defining problem	Complex, interpretative assessments that define what the problem is and reflect a medically informed conclusion	Diagnostic conclusion, etiological inference, prognostic judgment
Treatment goal	Decision to set a defined goal for treatment and thereby being more specific than giving advice	Quantitative or qualitative
Drug	Decision to start, refrain from, stop, alter or maintain a drug regimen	Start, stop, alter, maintain, refrain
Therapeutic procedure	Decision to intervene on a medical problem, plan, perform or refrain from therapeutic procedures	Start, stop, alter, maintain, refrain
Evaluating test result	Simple, normative assessments of clinical findings and tests	Positive, negative, ambiguous test results
Deferment	Decision to actively delay a decision or rejection to decide on a problem presented by a patient	Transfer responsibility, wait and see, change subject
Advice and precaution	Decision to give the patient advice or precaution, transferring responsibility for action to the patient	Advice or precaution
Legal/insurance related	Medical decision concerning to legal regulations or financial arrangements	Sick leave, drug refund, insurance, disability

Table 1: Overview of the ten medical decision categories defined in MedDec, together with a brief description of each category and representative examples. The Table is reproduced from (Elgaar et al., 2024).

Decision Type	**Sex**		**Race**						**Lng. Proficiency**	
	Male (n=259)	**Female** (n=192)	**White** (n=327)	**AA** (n=42)	**Hispanic** (n=25)	**Asian** (n=15)	**NH** (n=1)	**Other** (n=21)	**En** (n=260)	**Non-En** (n=45)
Defining Problem	39.2	38.8	39.5	37.5	38.0	36.4	30.9	38.6	38.7	39.2
Drug	26.0	25.1	25.7	24.4	25.0	27.5	19.1	27.0	26.1	25.6
Evaluation	12.9	13.6	12.6	16.6	13.3	12.7	25.5	12.8	13.1	13.9
Therapeutic proc.	12.2	12.4	12.4	12.5	11.7	13.2	10.6	12.2	12.0	12.0
Contact	4.9	5.2	5.0	4.6	6.0	5.4	8.5	4.3	4.8	5.1
Advice	3.4	3.5	3.5	3.2	4.2	3.3	0.0	3.9	3.9	3.0
Gathering info	0.8	0.9	0.8	0.7	1.2	1.3	5.3	0.9	0.9	0.6
Treatment goal	0.3	0.3	0.3	0.3	0.4	0.2	0.0	0.2	0.2	0.4
Deferment	0.2	0.2	0.2	0.2	0.2	0.0	0.0	0.1	0.2	0.2
Legal/Insurance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Count	33,054	24,235	41,666	5,684	3,264	1,737	94	3,078	37,026	6,295

Table 2: Percentage of annotated spans for each decision category across protected variables in MedDec. n is the number of the discharge summaries for each category. The last row shows the total count of decisions per variable.

2 MedExACT 2026 Task Description

The task is built on MedDec (Elgaar et al., 2024), which contains 451 discharge summaries, more than 54k sentences, and roughly 1.4M tokens annotated by domain experts. It reports token-level inter-annotator agreement of 0.74 Cohen's κ. The notes come from MIMIC-III, a de-identified critical care database distributed under controlled access (Johnson et al., 2016). MedExACT inherits that access model: participants obtain the training and validation data from PhysioNet, while the official gold labels for test set remain hidden.

Given a full discharge summary, systems must detect contiguous text spans that express medical decisions and assign each span one of nine DICTUM decision categories defined in Table 1: *Contact related, Gathering information, Defining problem, Treatment goal, Drug, Therapeutic procedure, Evaluating test result, Deferment, Advice and precaution*, or *None* for when no decision is present. Table 2 shows the distribution of the decision categories across protected variables in MedDec.

For MedExACT 2026, the released split contains 350 training summaries, 53 validation summaries, and the 48 hidden test summaries (Table 3). In addition to text and span annotations, the task package includes demographic attributes used for subgroup evaluation and phenotype metadata inherited from the MedDec release. For privacy reasons, access to the dataset requires approval for MIMIC-III under their data use agreement.

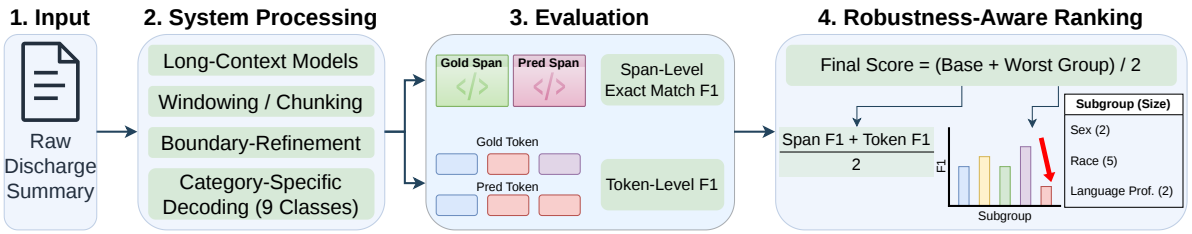


Figure 1: Overview of MedExACT 2026. (1) A raw ICU discharge summary from MIMIC-III is passed to the system. (2) Participating systems typically combine long-context encoders, windowing or chunking of long notes, boundary refinement, and category-specific decoding over the nine DICTUM decision classes to emit character-level span predictions. (3) Predictions are matched against hidden gold annotations at two granularities, span-level exact-match F1 (strict offset and label match) and token-level macro-F1, which are averaged into a per-note Base score. (4) The final leaderboard score is the mean of the overall Base score and the minimum Base score across nine subgroups defined by sex (2), race (5), and English proficiency (2), rewarding both accuracy and robustness.

As Figure 1 illustrates, the input to a submitted system is the raw text of a discharge summary taken from MedDec, and the output is a list of predictions, each represented by a start character offset, an end character offset, and a category ID. These predictions are matched against hidden gold annotations using a strict span-level evaluator together with a token-level evaluator, and the resulting Base score is combined with a worst-subgroup term to produce the final leaderboard score (Section 3). The shared task scores nine in-scope decision types. The underlying MedDec annotations follow DICTUM (Ofstad et al., 2016). Category 10 (Legal/insurance related) is ignored by the official shared-task evaluator.

Example: Each MedDec annotation includes annotator ID, discharge summary ID, and a list of decision spans with their text, DICTUM category, character offsets, and annotation ID:


```
{
 "annotator_id": "CALIML",
 "id": "10814_101543_52781",
 "annotations": [
  {
   "decision": "arrest with heart block",
   "category": "3: Defining problem",
   "start_offset": "526",
   "end_offset": "556",
   "annotation_id": "AC_005"
  },
  ...
  {
   "decision": "GU: Foley in place",
   "category": "6: Therapeutic procedure",
   "start_offset": "4000",
   "end_offset": "4018",
   "annotation_id": "AC_048"
  }
 ]
}
```


Split	#Notes	#Decisions	Access
Train	350	43,640	Released
Validation	53	7,044	Released
Test	48	6,537	Hidden

Table 3: MedExACT 2026 data split summary. The hidden test set is used only for official evaluation.

3 Evaluation

Submissions are assessed based on both performance and robustness across patient subgroups. Evaluation script is available on GitHub.[1]

Base Performance: is accessed by evaluating two core metrics: (i) *Span-F1*, which requires predicted spans to exactly match the gold spans (in both offset and label). The script performs a light word-boundary expansion and punctuation trimming before matching. (ii) *Token-F1*, computes macro-F1 per note after tokenizing the raw text with NLTK and then averages across notes. At token level, the evaluator excludes tokens that are covered by more than one gold span.

$$\text{Base} = (\text{SpanF1} + \text{TokenF1})/2$$

Robustness: Let $\mathcal{G}$ denote the set of subgroup buckets defined over sex, race, and language proficiency. The subgroup definitions used are Female/Male for sex, White/African American/Hispanic/Asian/Other for race, and English/Non-English for language. We compute the same Performance Base score separately for each of these nine subgroups and then calculate the minimum subgroup Base Score across all the

[1] github.com/CLU-UML/MedDec/blob/main/evaluate.py

Team	LLM	Ens	Cat	Long	Aug	Fair	Final
CUAMC (Baumgartner and Schilling, 2026)		✓				✓	0.596
LAMAR (Chiewhawan et al., 2026)	✓	✓			✓		0.594
Otter (Lowphansirikul and Ittichaiwong, 2026)		✓	✓		✓	✓	0.581
ELiRF-UPV (Ahuir et al., 2026)	✓		✓				0.581
MedMBZ (Elshehaby et al., 2026)		✓	✓	✓		✓	0.572
CASPAR (Tao et al., 2026)				✓			0.567
Kondadadi (Kondadadi, 2026)		✓		✓	✓		0.555
avishek		✓				✓	0.537
Fuyou Mao				✓		✓	0.533
ccccgo		✓		✓		✓	0.533
Aurum (Kumari et al., 2026)				✓	✓		0.525
dipika_nath		✓		✓	✓		0.514
TamuNLP		✓					0.512
Baseline				✓			0.511
CanSA (Alliheedi et al., 2026)	✓			✓			0.405
NoviceTrio	✓			✓			0.378

Table 4: Method matrix for the 15 system descriptions, sorted by descending final score, with the organizer Baseline (Elgaar et al., 2024) shaded for reference. LLM: LLM used in the prediction pipeline. ENS: multi-model ensemble or aggregation. CAT: category-specific decoding, routing, or reclassification. LONG: explicit long-document windowing or chunking. AUG: synthetic, pseudo-labeled, or augmented training data. FAIR: fairness-aware training, sampling, or model selection. Nearly all systems also use an encoder backbone and some form of boundary post-processing; those near-universal features are omitted.

subgroups. This metric rewards systems that optimize across all target subpopulations rather than optimizing only the global average; this explicitly incentivizes fairness-aware system design instead of treating fairness as a post-hoc diagnostic:

$$\text{WorstGroup} = \min_{g \in \mathcal{G}} \text{Base}_g$$

The final leaderboard ranking uses the following fairness-aware aggregate score:

$$\text{FinalScore} = (\text{Base} + \text{WorstGroup})/2$$

We consider the following two implementation details while evaluating: first, predictions must use character offsets into the exact raw text supplied to the evaluator. Any normalization of the whitespace, punctuation, or section structure can shift offsets and reduce scores; second, the shared task only scores category IDs 1–9. Gold spans or predictions outside that label space are ignored.

4 Task Resources and Public Baseline

In addition to the submission endpoint, the released repository includes official split files without text spans (the actual spans can be obtained from PhysioNet), a sample predictions JSON file, a data-cleaning utility for offset normalization, a statistics generator for subgroup evaluation, raw-text extraction scripts for MIMIC-III notes, and the official evaluator. The released bundle is also intended to support replication after the workshop period.

The public baseline fine-tunes RoBERTa for span tagging: tokens receive multiclass BIO-style labels over 512-token segments, and the resulting token labels are converted back to character-offset predictions for evaluation. The implementation also includes optional CRF decoding and random cropping of long notes during training. This baseline is meant to help participants verify preprocessing, training, and submission formatting before developing their own systems; it should not be treated as the definitive performance reference for the leaderboard. The original MedDec paper provides further baseline comparisons (Elgaar et al., 2024).

5 Participating Systems

The system-description file available to us contains 16 submissions from 15 organizations, and the leaderboard snapshot contains 37 scored runs (including the baseline model). Because the leaderboard file records submission usernames rather than canonical team names, our quantitative analy-

sis is run-centric, whereas the method analysis below uses the author-provided system descriptions when they are available.

Table 4 summarizes six design dimensions across the 15 described systems. Ensembles appear in 9 descriptions, explicit long-document windowing in 8, fairness-aware training or model selection in 6, data augmentation in 5, LLMs use in 4, and category-specific decoding or routing in 3. Nearly all systems also use an encoder backbone and some form of boundary post-processing; those near-universal features are omitted from the matrix.

The most discriminating column is category-specific processing. All three systems that route predictions through per-category decoders, reclassifiers, or routing branches—Otter (Lowphansirikul and Ittichaiwong, 2026) (category reclassifier), ELiRF-UPV (Ahuir et al., 2026) (independent CRFs per category with a section-category layer mixer), and MedMBZ (Elshehaby et al., 2026) (sparse category routing with span-length calibration)—were in the top five. This pattern suggests that a single set of decoding thresholds fits the nine heterogeneous DICTUM categories poorly and that category-aware specialization provides consistent gains. Ensembles and fairness-aware optimization also skew toward the top: four of the five highest-scoring systems use multi-model aggregation, and three explicitly optimize for the worst-group term in the official metric.

Explicit long-document windowing is common but does not predict high final scores on its own. The top two systems (CUAMC (Baumgartner and Schilling, 2026) and LAMAR (Chiewhawan et al., 2026)) do not flag windowing as a primary design choice, relying instead on architectures that handle length natively or on section-level decomposition. Among the four LLM-based systems, fine-tuned LLMs with ensemble aggregation (LAMAR) reach the top of the leaderboard, while prompt-based LLM pipelines (CanSA (Alliheedi et al., 2026), NoviceTrio) fall well below the median, confirming that supervision and structured decoding remain critical even when the base model is large.

6 Results

Table 5 reports the full test-phase leaderboard snapshot with all 37 submitted runs, using the organizer-provided display names listed in the paper. The top five entries are CUAMC, LAMAR, Otter, ELiRF-UPV, and MedMBZ. Across all 37 runs, the median final score is 0.533, while the mean of the top five final scores is 0.585. The organizers' baseline, shown as MedExACT Organizers, ranks 28th with a final score of 0.511. The gap between the median and the top runs shows that small modeling choices matter under this setup.

Submission	Final	Base	Worst	Span	Token
CUAMC	0.596	0.604	0.589	0.542	0.667
LAMAR	0.594	0.600	0.588	0.526	0.675
Otter	0.581	0.592	0.569	0.518	0.667
ELiRF-UPV	0.581	0.589	0.572	0.524	0.654
MedMBZ	0.572	0.585	0.560	0.490	0.680
NA	0.572	0.584	0.560	0.489	0.679
venus	0.572	0.585	0.558	0.490	0.680
fauna_rhea	0.571	0.584	0.558	0.490	0.678
jordanaskanov	0.570	0.582	0.557	0.489	0.675
CASPAR	0.567	0.573	0.560	0.506	0.641
NA	0.564	0.576	0.552	0.488	0.664
Kondadadi	0.555	0.577	0.534	0.482	0.673
fnkll	0.554	0.575	0.532	0.491	0.659
kotymoty	0.553	0.573	0.532	0.490	0.655
avishek	0.537	0.557	0.517	0.485	0.629
uijdsadada	0.533	0.541	0.525	0.430	0.652
ccccgo	0.533	0.541	0.525	0.430	0.652
csu-Medical	0.533	0.541	0.525	0.430	0.652
wubeining123	0.533	0.541	0.525	0.430	0.652
tekak_xo	0.529	0.549	0.508	0.466	0.632
rs-submission	0.529	0.549	0.508	0.466	0.632
popopop	0.528	0.535	0.521	0.438	0.632
Aurum	0.525	0.536	0.514	0.441	0.631
dbasudbasidad	0.515	0.523	0.508	0.412	0.633
dipika_nath	0.514	0.537	0.491	0.422	0.652
suxinqi	0.512	0.518	0.507	0.387	0.649
sususu123	0.512	0.517	0.506	0.385	0.649
Baseline	0.511	0.530	0.492	0.436	0.624
rm-rf-humans	0.510	0.544	0.476	0.461	0.627
nrgdedede	0.509	0.525	0.493	0.439	0.612
john_real	0.509	0.536	0.482	0.465	0.606
sususu1998	0.506	0.512	0.499	0.384	0.641
sususu666	0.477	0.483	0.470	0.349	0.617
CanSA	0.405	0.419	0.392	0.341	0.497
NoviceTrio	0.378	0.408	0.348	0.239	0.578
shahrin	0.264	0.274	0.255	0.215	0.334
mehulit22	0.031	0.047	0.015	0.006	0.088

Table 5: Full test-phase leaderboard snapshot with all submitted runs, sorted by descending final score of teams. The top five entries are marked with inline badge icons and subtle row shading.

The top systems outperform the baseline across almost all subgroups (Figure 2), however the strongest system is not uniformly best for every subgroup. The leaderboard also reveals a consistent trade-off between token overlap and exact span recovery (Figure 3). Among the five best final-score runs, CUAMC has the strongest span F1, while MedMBZ has the strongest token F1. Similar behavior appears further down the leaderboard: several runs reach token F1 in the 0.67–0.68 range while remaining below 0.53 final score because exact

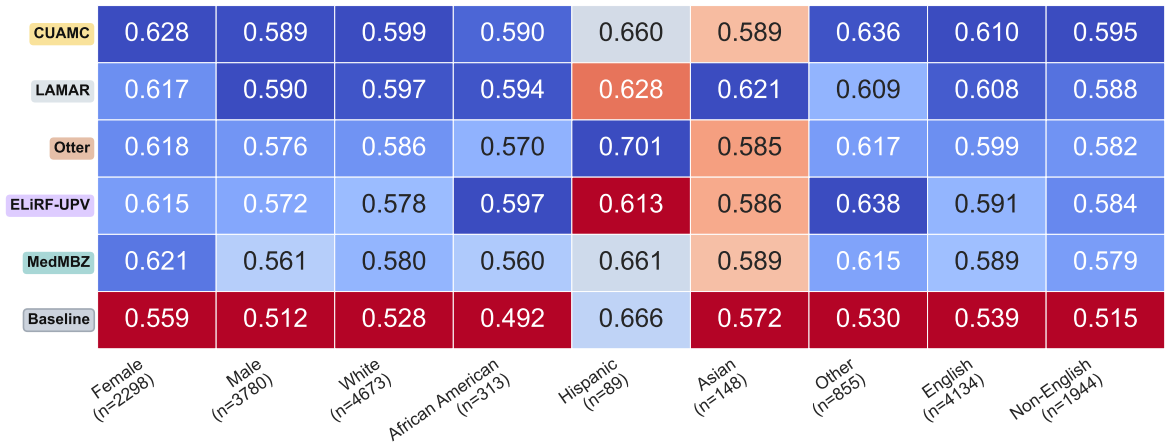


Figure 2: Subgroup Base scores for the top five MedExACT 2026 systems and the organizer baseline. Cell values show raw subgroup Base scores; colors are normalized within each subgroup to emphasize relative robustness patterns across sex, race, and English-proficiency groups.

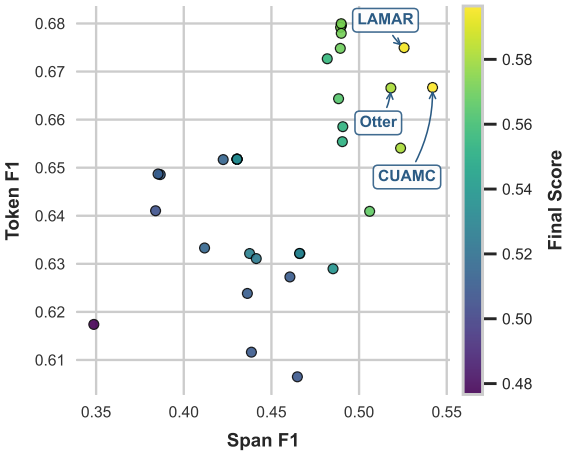


Figure 3: Scatter plot of Span F1 versus Token F1 across MedExACT 2026 leaderboard submissions. The top three systems (`CUAMC`, `LAMAR`, and `Otter`) are highlighted. High token-level overlap frequently fails to yield exact span boundary recovery.

span boundaries and subgroup robustness remain limiting factors. High token F1 with lower span F1 is consistent with coarse localization followed by boundary errors under the strict span matcher.

The fairness-aware objective also changes the interpretation of the ranking. For the top four runs, the worst-group score trails the base score by roughly 0.012–0.023 absolute points. These gaps are small enough to indicate deliberate robustness tuning, yet large enough to affect final ordering when systems optimize average accuracy only. Participant descriptions confirm that many teams treated the official final score as the primary development objective rather than as a secondary reporting metric.

7 Discussion

We summarize three takeaways from the participant reports and leaderboard snapshot:

Long-Context Modeling is a Core Requirement: Handling long discharge summaries is a prerequisite for this task. Given the length and density of ICU discharge summaries, most competitive systems used overlapping windows, chunk-level processing, or section-aware decomposition, including large encoder ensembles, Clinical-Longformer fine-tuning, and retrieval-based prompting pipelines (Kondadadi, 2026; Kumari et al., 2026; Alliheedi et al., 2026).

Exact Span Recovery is the Primary Bottleneck: Strong token F1 values show that systems can often detect the right region of text and recover the correct semantic type. The harder step is to return the precise character span expected by the official evaluator. Boundary refinement, CRF decoding, overlap-aware aggregation, and second-pass calibration all target this failure mode, so similar modules appear across otherwise different model families (Tao et al., 2026; Chiewhawan et al., 2026; Elshehaby et al., 2026).

Robustness-Aware System Design: Because the final ranking combines average performance with worst-group performance, teams invested in oversampling, subgroup-aware loss design, adversarial training, and development-time checkpoint selection under the official score (Baumgartner and Schilling, 2026; Lowphansirikul and Ittichaiwong, 2026; Elshehaby et al., 2026). For future clinically

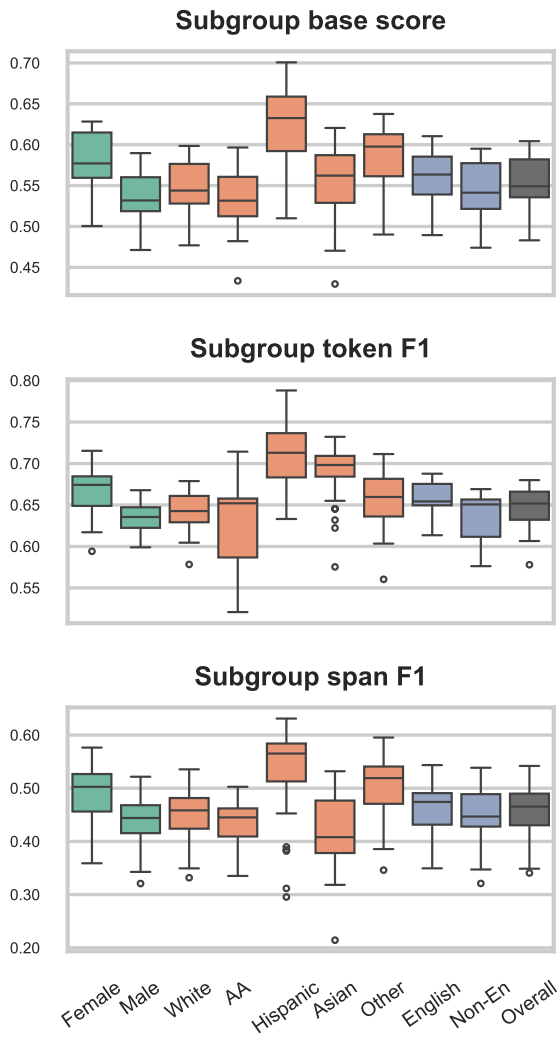


Figure 4: Distribution of Base score, Token F1, and Span F1 across submissions by subgroup and overall performance. Boxplots summarize variation across runs for sex, race, language, and overall scores, showing that subgroup differences persist across all categories.

grounded shared tasks, this example suggests that subgroup robustness should be built into the primary ranking rule instead of a post hoc diagnostic.

The Role of LLMs in Decision Extraction: The submitted systems also clarify the current role of LLMs in this space. Prior work has shown that large language models can be effective few-shot clinical information extractors in some settings (Agrawal et al., 2022). In MedExACT, however, long discharge summaries and strict character-offset evaluation seem to favor LLMs as augmentation, reranking, or repair modules instead of as standalone span extractors (Chiewhawan et al., 2026; Alliheedi et al., 2026). This matches the low zero-shot and one-shot exact-match results reported in the MedDec paper (Elgaar et al., 2024).

Toward Generalization: A natural next step for this shared task is to move beyond in-domain robustness and evaluate whether systems can generalize across clinically distinct settings under explicit domain shift. A future version of MedExACT could frame this as phenotype-family transfer, where models are trained without access to one target phenotype family and then tested on discharge summaries from that unseen family. This new setting measures whether systems learn decision semantics instead of phenotype-specific lexical or discourse patterns. We can annotate discharge summaries with their corresponding phenotypes, group phenotypes into clinically meaningful families, and reports leave-one-family-out results. We expect this setting to be feasible and result in substantially lower than in-domain performance due to the difficulty of transfer to unseen clinical contexts. This direction would complement the current robustness-aware evaluation by testing not only subgroup stability, but also resilience to shifts in patient populations and clinical conditions, and could motivate new methods for domain generalization, transfer learning, and clinically grounded robustness in medical decision extraction.

Understanding Clinical Decisions: This task can help us better understand underlying clinical decision-making by transforming unstructured clinical notes into structured representations of the decisions clinicians make. This enables large-scale analysis of how clinical reasoning is recorded across patients, conditions, and care settings. Such structured evidence could also provide better clinical decision support by making prior decisions easier to review, compare, and audit; helping clinicians recognize patterns in diagnosis and treatment; identifying inconsistencies in documentation (Amiri et al., 2024) and blind spots of extraction models (Elgaar and Amiri, 2026); and inform decision-support tools grounded in real clinical narratives.

8 Conclusion

MedExACT 2026 focuses on robust medical decision extraction over long discharge summaries. The submitted systems mainly use long-context encoder models, ensembles, boundary-aware decoding, and explicit optimization for the official worst-group-adjusted score. The task establishes a benchmark for medical decision extraction and shows how robustness-aware evaluation can influence system design in clinical NLP.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2021. Detect and classify – joint span detection and classification for health outcomes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8709–8721, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vicent Ahuir, Lluís Felip Hurtado, and María José Castro-Bleda. 2026. ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mohammed Alliheedi, Robert E. Mercer, Anemily V. Machina, Sudipta Singha Roy, Yetian Wang, and Xindi Wang. 2026. CanSA at MedExACT@ACL 2026: Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Hadi Amiri, Nidhi Vakil, Mohamed Elgaar, Jiali Cheng, Mitra Mohtarami, Adrian Wong, Mehrnaz Sadrolashrafi, and Leo A Celi. 2024. Analysis of race, sex, and language proficiency disparities in documented medical decisions. *medRxiv*, pages 2024–07.

William Baumgartner and Lisa M. Schilling. 2026. CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Monrada Chiewhawan, Keetawan Limaroon, and Titipat Achakulvisut. 2026. LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mohamed Elgaar and Hadi Amiri. 2026. Linguistic blind spots in clinical decision extraction. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 46–54, Rabat, Morocco. Association for Computational Linguistics.

Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, and Leo Anthony Celi. 2025. MedDecXtract: A clinician-support system for extracting, visualizing, and annotating medical decisions in clinical narratives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 481–489, Vienna, Austria. Association for Computational Linguistics.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. MedDec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.

Ahmed Elshehaby, Mohamed Abdalla, and Youssef MK Mohamed. 2026. Sparse Category Routing and Fairness-Aware Optimization for Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

Rishik K. Kondadadi. 2026. Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Jyoti Kumari, Vinay Babu Ulli, and Anindita Mondal. 2026. Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Longformer Fine-Tuning for Medical Decision Extraction. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, Florence T. Bourgeois, and Adam G. Dunn. 2021. Sent2Span: Span detection for PICO extraction in the biomedical text without span annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1705–1715, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lalita Lowphansirikul and Piyalitt Ittichaiwong. 2026. Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. Association for Computational Linguistics.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

Eirik H. Ofstad, Jan C. Frich, Edvin Schei, Richard M. Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ Open*, 6(2):e010098.

Jing Tao, Amir Eskandari, and Farhana Zulkernine. 2026. CASPAR: A Context-Aware Span Refinement Approach for Decision Support. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*. Association for Computational Linguistics.

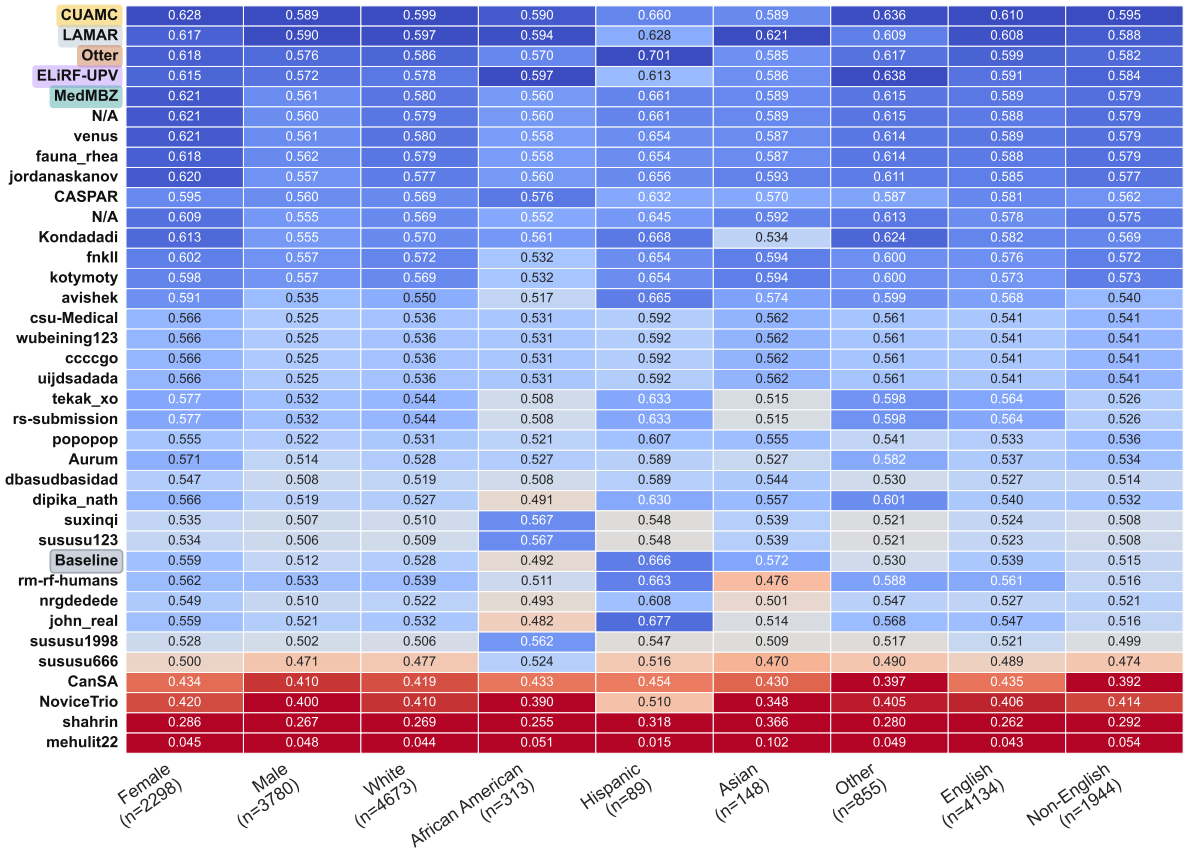


Figure 5: Full subgroup Base-score heatmap for all MedExACT 2026 leaderboard submissions. Rows are ordered by final leaderboard score; columns correspond to the nine sex, race, and English-proficiency subgroups used in the official robustness metric.

A Full Subgroup Leaderboard

Figure 5 shows the full subgroup Base-score heatmap for all MedExACT 2026 leaderboard submissions. Rows are ordered by final leaderboard score; columns correspond to the nine sex, race, and English-proficiency subgroups used in the official robustness metric. Three patterns stand out. First, the highest-ranked systems are strong across nearly all columns rather than winning through a single favorable subgroup; their Base scores mostly stay in the 0.58–0.66 range. Second, no system dominates every subgroup: `CUAMC` has the best English and Non-English scores, `LAMAR` is strongest on Asian patients, `Otter` is strongest on Hispanic patients, and `ELiRF-UPV` is strongest on African American and Other race groups. Third, subgroup robustness remains a real source of variation. Several mid-ranked runs have one or two noticeably weak columns, and the public baseline is illustrative: it is competitive on Hispanic patients but drops on African American, Male, and Non-English patients. The lowest-ranked systems are weak across all subgroups, which suggests that the robustness metric mainly separates otherwise capable systems rather than rescuing globally poor extractors.

Divide-Prompt-Refine: a Training-Free, Structure-Aware Framework for Biomedical Abstract Generation

Sylvey Lin[1], Joe Menke[1], Shufan Ming [1], Dongin Nam[1], Neil Smalheiser[1,2], Halil Kilicoglu[1],

[1] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL
[2] Department of Psychiatry, University of Illinois College of Medicine, Chicago, IL

Correspondence: yuhsinl2@illinois.edu

Abstract

Biomedical abstracts play a critical role in downstream NLP applications, such as information retrieval, biocuration, and biomedical knowledge discovery. However, a non-trivial number of biomedical articles do not have abstracts, diminishing the utility of these articles for downstream tasks. We propose DPR-BAG (Divide, Prompt, and Refine for Biomedical Abstract Generation), a training-free, zero-shot framework that generates coherent and factually grounded abstracts for biomedical articles with full text but no abstract. DPR-BAG decomposes full-text documents into structured rhetorical facets following the Background-Objective-Methods-Results-Conclusions (BOMRC) schema, performs parallel LLM-based summarization for each facet, and applies a final refinement stage to restore global discourse coherence. On PMC-MAD, a distribution-aligned dataset of 46,309 biomedical articles, DPR-BAG improves abstractive novelty over strong extractive and fine-tuned baselines, while maintaining factual consistency. Our ablation study reveals a counterintuitive finding: increasing prompt complexity or explicitly injecting entity-level guidance can degrade factual alignment, highlighting the importance of controlled prompting strategies. These findings underscore the potential of training-free, structure-aware frameworks for scalable biomedical abstract generation in low-resource settings. Our data and code are available at https://huggingface.co/datasets/pmc-mad/PMC-MAD and https://github.com/ScienceNLP-Lab/MultiTagger-v2/tree/main/DPR-BAG.

1 Introduction

Many biomedical NLP tasks rely heavily on abstracts, due to their accessibility and information density. Abstracts provide an author-written summary of core scientific findings, making them a useful proxy for full-text articles in downstream applications. For example, Luo et al. (2022) showed that pre-training tasks designed around the title-abstract structure improve biomedical information retrieval; Wiegers et al. (2025) used abstracts as an initial data source for biocuration. Beyond content, the structured organization of abstracts also benefits downstream tasks: Ueda et al. (2021) leverage abstract-level structure to refine retrieval; PubMedQA (Jin et al., 2019) is based on structured abstracts to support high-fidelity biomedical knowledge discovery. Moreover, abstracts alone can serve as a stronger training signal than full texts in some settings (Gu et al., 2021).

However, the absence of abstracts in a significant portion of biomedical articles creates a bottleneck for these tasks. As of April 2026, 11,603,796 out of 40,414,072 (~29%) PubMed articles were missing abstracts, with this volume continuing to rise despite a declining overall proportion, driven by the growth of publication types such as case reports, editorials, and letters. These publication types carry substantial scientific value. For instance, Gurulingappa et al. (2012) and Fan et al. (2020) utilized case reports for adverse drug event detection; Magnet and Carnet (2006) and Nuzzo (2021) assessed letters to characterize post-publication scientific discourse, including patterns of critique, rhetorical features of disagreement, and trends in authorship; and Waaijer et al. (2011) and Ioannidis and Schippers (2025) analyzed editorials to study how journals shape scientific discourse, including the distribution of topics, the framing of policy issues, and the presence of systematic biases.

The absence of abstracts in these articles motivates the Biomedical Abstract Generation (BAG) task, which aims to automatically generate abstracts from full-text biomedical articles. Although related to standard document summarization, BAG differs in important ways. It must adhere to scientific reporting conventions, including structured

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 770–790
July 3-4, 2026 ©2026 Association for Computational Linguistics

presentation of methods, results, and conclusions, while preserving fine-grained biomedical entities, quantitative findings, and explicit argumentative relationships that are often critical for scientific interpretation. Early BAG work by Chachra et al. (2016) utilized extractive sentence selection, which can lead to fragmented coherence and poor lexical flow. Moreover, because full-length biomedical articles often exceed the context limits of standard models, BAG is inherently a long-context task, making it vulnerable to extractive bias and factual fidelity issues. For example, Wang et al. (2025) demonstrates that even state-of-the-art models like GPT-4 suffer from hallucinations and information omission when extracting from non-decomposed scientific full texts, emphasizing the inherent fidelity risks in long-document processing required for BAG task. Beyond fidelity risks, recent analysis also reveals that when forced to process complex long full texts, even specialized models like LongT5 exhibit a strong extractive bias, relying on simple heuristics to copy verbatim snippets rather than synthesizing information (Chernyshev and Dobrov, 2024). As a result, these models can suffer from the same core issue as traditional extractive summarizers: they produce fragmented text that lacks the natural flow and cohesion of human-written summaries (Giarelis et al., 2023).

To address these limitations, we propose the Divide, Prompt and Refine Biomedical Article Generation (DPR-BAG) framework. Drawing on prior work showing that divide-and-conquer decomposition reduces intermediate errors in LLMs (Zhang et al., 2025), DPR-BAG decomposes full-text articles along their rhetorical structure, performs parallel summarization on each resulting facet, and applies a modular refinement stage to reconcile fragmented outputs and restore discourse coherence. We target six rhetorical facets: Background, Objectives, Methods, Results, Conclusions (BOMRC), and Others. BOMRC is adopted as it represents the standard discourse structure validated in the PubMed 200k RCT dataset (Dernoncourt and Lee, 2017), while the “Others" facet retains any unclassified content. Using this design, we focus on three research questions:

1. Can we develop a training-free approach for the BAG task?
2. Does structure-aware decomposition of full-text articles improve the quality of generated abstracts compared to naive prompting?
3. To what extent does increasing prompting complexity (from detailed instructions to entity guidance) improve generation quality?

Our main contributions are as follows:

1. We propose DPR-BAG, a training-free, structure-aware method for BAG.
2. We release a dataset of more than 46K biomedical full-text publications for BAG task.
3. We compare DPR-BAG to strong extractive and abstractive baselines.
4. We systematically evaluate the effect of various prompting and splitting strategies as well as entity guidance within DPR-BAG.

2 Dataset

We constructed a BAG dataset based on PubMed publications from 1987 to 2023. To ensure a representative sample, we first calculated the publication type (PT) distribution of articles lacking abstracts using PT queries adapted from prior work (Menke et al., 2024). We then performed stratified sampling based on this distribution to retrieve 130,000 candidate XML files from the PubMed Central (PMC) Open Access subset, ensuring that the sampled articles reflect the publication type distribution of abstract-less PubMed records. For data processing, we adapted the extraction pipeline from the Longsummarization framework (Cohan et al., 2018) to parse structured sections and abstracts from the raw XML files. After filtering out records that were unparseable or lacked extractable abstracts, the final dataset, hereafter referred to as PMC-MAD (Missing-Abstract Distribution-aligned PMC), consists of 46,309 articles.

3 Methods

DPR-BAG follows a modular pipeline designed to generate structure-aware abstracts from biomedical full-text articles (Figure 1). The process begins by decomposing the document into five distinct facets based on BOMRC, plus an "Others" facet for unclassified content. For each facet, we perform parallel LLM-based summarization, which can optionally be augmented with entity guidance extension. The resulting sectional summaries are then concatenated and passed to a final LLM-based refinement stage to restore discourse coherence. DPR-BAG

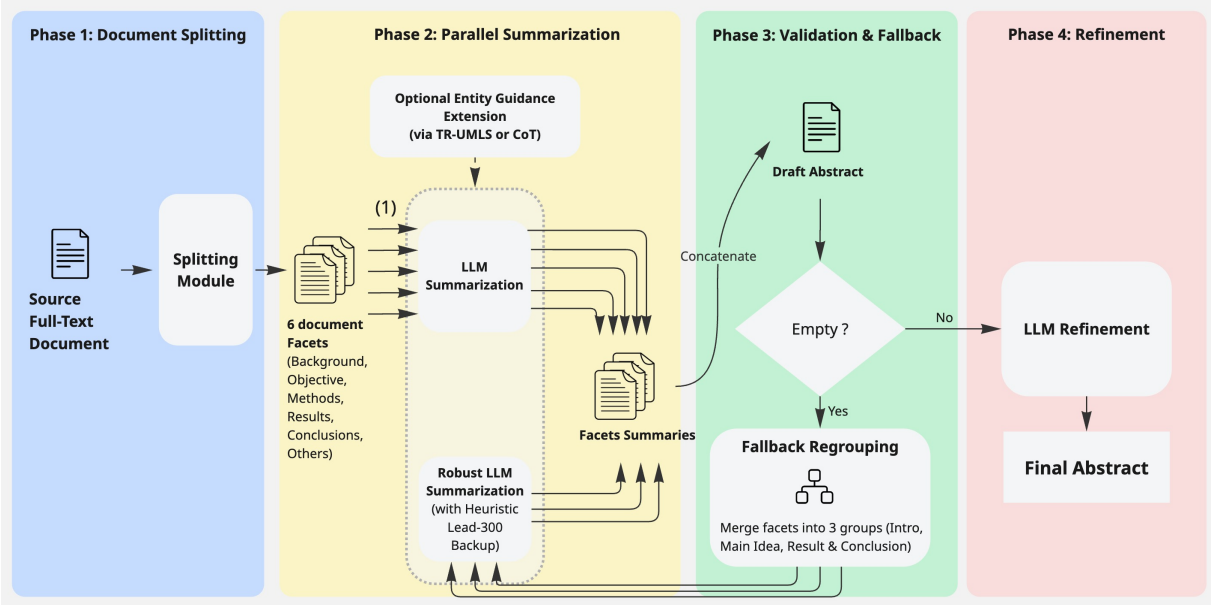


Figure 1: Overview of the DPR-BAG framework for biomedical abstract generation.

requires no task-specific training or fine-tuning; all components operate in a zero-shot manner using pre-trained, off-the-shelf models.

3.1 Task Formulation

Given a full-text biomedical article $D = (p_1, p_2, \ldots, p_n)$, where each p_i denotes a paragraph, the goal is to generate an abstract A that covers a predefined set of rhetorical facets $\mathcal{F} = \{\text{Background, Objectives, Methods, Results, Conclusions, Others}\}$ (BOMRC+).

We reformulate this as a facet-conditioned summarization problem, where the model generates content for each rhetorical facet separately. This decomposition allows the model to address each rhetorical component independently and capture discourse structure. Specifically, the document is partitioned into $K = 6$ facet-specific sub-documents $\{D_{f_k}\}_{k=1}^{K}$, where each D_{f_k} aggregates paragraphs rhetorically aligned with facet f_k. Each sub-document is independently summarized to produce a facet summary $\hat{a}_{f_k}$, and the concatenation $\hat{A} = \bigoplus_{k=1}^{K} \hat{a}_{f_k}$ is subsequently refined into the final abstract $R(\hat{A}) = A$. Facets absent from the source document yield empty strings.

3.2 Document Splitting

To segment full-text documents into rhetorically coherent texts, we use LLM-SSC (Lan et al., 2024), an LLM-based sequential sentence classification framework that assigns rhetorical labels (BOMRC) to sentences using in-context learning (performance details in Appendix E). While the model was trained on structured abstracts, we assume that the underlying rhetorical intent (e.g., methodological description vs. result reporting) remains consistent within full-text paragraphs. Specifically, we leverage the role of the first sentence of the paragraph as a topic sentence that typically encapsulates the paragraph's functional purpose. We assign the label of the first sentence of each paragraph as the global label for that paragraph, subsequently concatenating all paragraphs with matching labels to form the input document facets. We refer to this approach as the **First Sentence Labeling (FS)** strategy, and empirically validate it against naive splitting (NS) and section-header (SH) ablation variants below.

Naive Splitting Approach (NS): This approach distributes paragraphs into six segments, aiming for an approximately even distribution while maintaining paragraph integrity. Including this baseline allows us to assess whether a semantic-aware division (e.g., LLM-SSC) offers advantages over a purely structural, length-based partition.

Section-header Normalization (SH): This strategy serves as a coarse-grained semantic baseline. Utilizing the Transformer model developed in Lin et al. (2025), this approach categorizes existing section headers into standard BOMRC categories and concatenates paragraphs within the same facet. This comparison helps determine if the fine-grained, sentence-level classification used in LLM-SSC provides additional utility beyond

simple section-level organization.

3.3 Parallel Summarization

After the documents are divided into six document facets (BOMRC and Others), each is input into the LLM to generate a corresponding facet summary; facets that are not present in the source document are represented as empty summaries. These summaries are subsequently concatenated to form the draft abstract. The following subsections detail the prompting strategies and optional entity guidance extensions used during summarization.

3.3.1 Prompting Strategies

To investigate the effect of prompt complexity on generation quality, we adopt a **Basic Concise (BC)** prompting strategy as the baseline, and ablate prompting complexity by evaluating two more elaborate variants, **Detailed Instruction (DI)** and **Structural Instruction (SI)**. Full prompt templates are described in Appendix A.

Basic Concise Prompting (BC): BC is a minimal prompting strategy with coarse-grained focal points for each rhetorical facet (e.g., directing the model to "prioritize key findings and data" for the Results section) without further elaboration or explicit formatting structure.

Detailed Instruction Prompt (DI): DI is a more detailed prompting strategy modeled after the abstract submission guidelines of JMIR Publications[1] whose five-part BOMRC structured guideline aligns with the target rhetorical categories that DPR-BAG uses. By shifting the LLM persona to a "biomedical synthesis assistant," this prompt aims to enforce the extraction of granular details and mandates the inclusion of specific research designs, sample sizes, response rates, and statistical metrics (such as p-values and confidence intervals) to ensure adherence to multifaceted reporting standards.

Structural Instruction Prompt (SI): SI extends the basic prompt by introducing an explicit structural schema using Markdown formatting strategy, inspired by He et al. (2024). Compared to DI, which focuses on detailed content guidance, SI organizes the prompt into a more structured format, which aims to improve instruction adherence.

[1] https://support.jmir.org/hc/en-us/articles/37982552280987-Submitting-Your-Manuscript-to-JMIR-Publications-A-Guide-for-Authors

3.3.2 Entity Guidance

We additionally introduce an optional knowledge-grounding extension to further enhance semantic fidelity during parallel summarization. This component extracts key biomedical entities to guide the LLM in summarizing each facet. We consider two instantiations of this extension, TR-UMLS and CoT, detailed below.

TextRank and UMLS normalization (TR-UMLS): We extract key phrases from each facet using TextRank, a graph-based unsupervised method that requires no additional training and is well-suited to our zero-shot setting, and link them to UMLS (Bodenreider, 2004) concepts using scispaCy's UMLS entity linker (Neumann et al., 2019), retaining only phrases with valid UMLS mappings. When multiple phrases map to the same UMLS concept, they are grouped together and represented by a single term to avoid redundant anchoring. The top-n entities, ranked by TextRank centrality, are incorporated into the summarization prompt as anchor terms to guide the model toward the most structurally significant medical information in the text. We ablate $n \in \{5, 10\}$ in Section 5.2.

Chain-of-Thought (CoT): As an alternative, we employ a two-stage prompting strategy within the LLM summarization module when the extension is activated. In the first stage, we prompt the model to list the important entities of the facet, and in the second stage, the model is prompted to synthesize the information based on the facet and the entities it listed in the first stage.

We ablate these two entity guidance strategies independently. TR-UMLS pairs with DI, while CoT pairs with SI, whose structured format makes it well-suited for CoT's two-stage reasoning.

3.4 Validation and Fallback

To ensure the summarization pipeline robustness, we implement a validation and fallback mechanism. If all six facet summaries are empty due to insufficient context or LLM output format violations, the facets are regrouped into three broader categories (Intro, Main Idea, and Results & Conclusions) and sent back to the parallel summarization module. If regrouping fails, a first 300 characters (Lead-300) heuristic backup is used to guarantee non-empty output (details in Appendix B).

3.5 Refinement

Upon the formation of the draft abstract, we prompt the LLM to perform a global refinement to ensure structural coherence and stylistic consistency. This final processing step synthesizes the concatenated facets into a unified Final Abstract (the refinement prompt is detailed in Appendix C).

4 Experimental Setup

DPR-BAG is implemented with `Llama-3.2:3B` deployed via Ollama in instruction-tuned mode. Hardware details, token-limit constraints, and fine-tuning hyperparameters are provided in Appendix D.

4.1 Baseline Models

To establish robust baselines for our framework, we compared our approach against several standard long-document summarization models. We utilized two off-the-shelf variants of the Longformer Encoder-Decoder (LED) architecture (Beltagy et al., 2020)—pretrained on arXiv[2] and PubMed[3], respectively—to evaluate their zero-shot transferability to our task. Additionally, we included an off-the-shelf LongT5 model (Guo et al., 2022) pretrained on PubMed[4] to broaden our baseline comparisons across different architectures. Finally, to ensure maximal adaptation to our corpus, we evaluated a supervised fine-tuned version of the PubMed-pretrained LED and LongT5 using our 80% training split, using the 10% validation split for early stopping. All evaluations were performed on the remaining 10% (test split).

4.2 Evaluation Metrics

To assess the generated abstracts, we employ a multi-dimensional suite of metrics, with particular emphasis on abstractiveness and factuality. Detailed formulations and implementation details of each metric are provided in Appendix I. Paired bootstrap significance tests for the main comparisons are provided in Appendix J.

Abstractiveness: **Bigram and trigram novelty** measure the proportion of tokens absent from the source text, serving as a proxy for abstractive synthesis. We additionally adopt **Density** from Newsroom (Grusky et al., 2018), which quantifies the length of verbatim copying from the source, offering a complementary view of the model's extractive behavior. For both, smaller absolute deviations from the human-written reference indicate closer stylistic alignment.

Factuality: Given that our target abstracts are expected to be highly abstractive, factuality metrics must remain reliable under heavy paraphrasing. We therefore adopt **AlignScore** (Zha et al., 2023) as our primary factuality measure, as it evaluates alignment against the source full-text across a broad range of dimensions (notably paraphrasing), making it more robust than purely entailment-based alternatives. We additionally report **SummaC** (Laban et al., 2022) and **MiniCheck** (Tang et al., 2024), both NLI-based metrics, as cross-checks.

Semantic Alignment: We adopt **BERTScore** (Zhang et al., 2020) to evaluate semantic similarity between generated abstracts and the reference abstracts via contextual embeddings. To further assess discourse coherence, we adopt **DiscoScore** (Zhao et al., 2023), reporting DS_SENT_NN (sentence-level structural alignment) and DS_FOCUS_NN (shared noun semantic alignment).

Supporting Metrics: We additionally adopt **ROUGE-L** (Lin, 2004) to measure n-gram overlap with the reference. **UMLS Recall** quantifies the proportion of UMLS (Bodenreider, 2004) concepts from the original abstract retained in the generated output. **Coverage** and **Compression** from Newsroom (Grusky et al., 2018) serve as complementary metrics to measure how source content is preserved and condensed. For these two metrics, as well, smaller absolute deviations from the human-written reference are preferred.

5 Results

5.1 RQ1: Effectiveness of the Training-Free Approach

We first evaluate whether the proposed training-free approach can achieve competitive performance on the BAG task. As shown in Table 1 and Table 2, DPR-BAG (BC prompt, no entity guidance extension) achieves competitive performance against fine-tuned baselines. It generates abstracts that

[2] https://huggingface.co/allenai/led-large-16384-arxiv

[3] https://huggingface.co/patrickvonplaten/led-large-16384-pubmed

[4] https://huggingface.co/Stancld/longt5-tglobal-large-16384-pubmed-3k_steps

Model	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.495	0.663	6.650	–	–	–
LED-Arxiv (Base)	0.140 (-0.355)	0.222 (-0.441)	22.698 (+16.048)	0.720	0.856	**0.735**
LED-Pubmed (Base)	0.167 (-0.328)	0.261 (-0.402)	19.233 (+12.583)	0.667	0.845	0.693
LED-Pubmed (FT)	0.309 (-0.187)	0.453 (-0.210)	11.369 (+4.719)	0.511	0.653	0.503
LongT5 (Base)	0.174 (-0.321)	0.280 (-0.383)	15.492 (+8.842)	0.668	0.875	0.711
LongT5 (FT)	0.252 (-0.243)	0.379 (-0.284)	12.989 (+6.339)	0.540	0.725	0.565
DPR-BAG	**0.397 (-0.098)**	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.570

Table 1: Performance comparison on primary evaluation dimensions. For Abstractiveness metrics, parenthetical values indicate the difference from the original abstract; best = smallest absolute deviation. Factuality scores are computed against the source full-text; higher is better. Best results bolded. (Bi-g: Bigram novelty; Tri-g: Trigram novelty; Dens.: Density; AS: AlignScore; MC: MiniCheck)

Model	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.743	23.827
LED-Arxiv (Base)	0.635	0.907	0.702	0.234	0.340	0.950 (+0.217)	26.916 (+3.089)
LED-Pubmed (Base)	0.652	**0.908**	**0.670**	0.262	0.354	0.944 (+0.201)	27.387 (+3.560)
LED-Pubmed (FT)	0.634	0.868	0.791	0.264	**0.364**	**0.868 (+0.125)**	**20.901 (-2.926)**
LongT5 (Base)	**0.664**	0.876	0.861	0.274	0.312	0.948 (+0.205)	46.449 (+22.622)
LongT5 (FT)	0.650	0.895	0.713	**0.277**	0.361	0.911 (+0.168)	27.002 (+3.175)
DPR-BAG	0.617	0.868	0.828	0.183	0.266	0.894 (+0.151)	50.037 (+26.210)

Table 2: Performance comparison on semantic alignment and supporting metrics. For Coverage and Compression, parenthetical values indicate the difference from the original abstract; best = smallest absolute deviation. BS, SENT, R-L, and U-R: higher is better. FOCUS: lower is better. Best results bolded. (BS: BERTScore F1; SENT: DS_SENT_NN; FOCUS: DS_FOCUS_NN; R-L: ROUGE-L; U-R: UMLS Recall; Cov.: Coverage; Comp.: Compression)

more closely resemble human-written abstracts (more abstractive), while maintaining factual consistency with the full text (Table 1).

DPR-BAG outperforms the fine-tuned baselines in both abstractiveness and factuality (all paired bootstrap $p < 0.001$ on Bigram and Trigram novelty, Density, AlignScore, and MiniCheck; Appendix J.2), while the fine-tuned baselines themselves perform better than other baselines in abstractiveness but their generations are less factual.

DPR-BAG yields mostly lower scores for semantic alignment and other supporting metrics compared to the baselines, reflecting a known bias in these metrics toward extractive outputs (Table 2). Baseline models exhibit high density and low novelty relative to human-written abstracts (Table 1), indicating extractive behavior that likely inflates their scores. At the same time, DPR-BAG yields the highest compression rate (50.037), potentially leading to over-simplification of key information. We provide qualitative examples of generated abstracts in Appendix K.

5.2 RQ2: Impact of Structure-Aware Decomposition

We next evaluate whether structure-aware decomposition improves generation quality compared to naive prompting. As shown in Table 3, FS achieves the best overall performance. Compared to NS, FS achieves significantly higher AlignScore (paired diff = +0.006, $p < 0.05$), with no significant difference on MiniCheck or SummaC. Both approaches have similar abstractiveness, with NS achieving scores marginally closer to the human-written abstracts on Trigram novelty. Comparing FS with SH further underscores the necessity of fine-grained local context: SH significantly degrades AlignScore (-0.126), MiniCheck (-0.099), and SummaC (-0.141), all $p < 0.001$, indicating that broad semantic boundaries provided by section headers fail to provide sufficient contextual anchoring for faithful generation.

5.3 RQ3: Impact of Prompting Strategy and Entity Guidance

Next, we examine the effect of increasing prompt complexity.

Prompting Strategy: While DI and SI yield marginal but significant improvements in BERTScore ($p < 0.001$), both suffer from degradation in AlignScore, MiniCheck, SummaC, and UMLS Recall compared to BC (all $p < 0.001$; Table 4). This suggests that dense, multi-faceted

Splitting	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
FS	0.605 (-0.058)	**4.690 (-1.960)**	**0.762**	0.890	0.617	**0.828**	**0.266**
NS	**0.613 (-0.050)**	4.508 (-2.142)	0.756	**0.892**	0.615	0.856	0.262
SH	0.746 (+0.083)	2.885 (-3.765)	0.636	0.791	**0.636**	0.920	0.229

Table 3: Ablation on document splitting strategies. Splitting strategies are Naive Splitting (NS), Section Header Normalization (SH), and First Sentence Labeling (FS). (Abbreviations as in Table 1 and Table 2)

Prompt	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
DI	0.725 (+0.063)	3.131 (-3.519)	0.642	0.806	**0.638**	0.885	0.247
SI	0.736 (+0.073)	3.049 (-3.601)	0.630	0.805	0.636	0.836	0.251

Table 4: Ablation on summarization prompt strategies. Prompt variants are Basic Concise (BC), Detailed Instruction (DI), and Structural Instruction (SI). (Abbreviations as in Table 1 and Table 2)

instructions might have introduced a distraction effect, where the model struggles to simultaneously satisfy formatting constraints and maintain source-grounded factual alignment, motivating the need for explicit reasoning pathways and external grounding.

Entity Guidance: To assess whether external grounding can recover the factual consistency degradation observed in DI and SI, we ablate two entity guidance approaches: TR-UMLS integrated into DI, and CoT integrated into SI, with BC (no entity guidance) serving as the reference baseline.

As shown in Table 5, applying TR-UMLS to DI does not improve the overall performance. Top-5 and top-10 variants show no significant change in AlignScore relative to the base DI prompt, and both remain significantly below the BC baseline ($p < 0.001$). DS-Focus score also increases under top-5 entities ($p < 0.01$), suggesting that explicit entity conditioning might cause the model to overprioritize the provided terms, exacerbating the distraction effect rather than serving as an effective grounding mechanism.

Table 6 shows that integrating CoT into the SI prompt improves semantic alignment with the original abstract compared to BC (BERTScore +0.015, $p < 0.001$) while yielding overall higher abstractiveness and lower factuality (AlignScore -0.166, MiniCheck -0.098, SummaC -0.158, all $p < 0.001$). This shows that CoT encourages more abstractive and novel phrasing, at the expense of reducing the factual overlap with the source.

6 Discussion

Our results show that the proposed training-free approach generates abstracts that more closely resemble human-written abstracts in terms of abstractiveness and factuality. Notably, models fine-tuned on PMC-MAD do not match this performance on these dimensions. However, these fine-tuned models achieve slightly better semantic alignment, lexical/semantic overlap, and compression than DPR-BAG. Since the baseline models do not explicitly model rhetorical structure, we attribute the stronger performance of DPR-BAG on abstractiveness and factuality to its structure-aware design. Specifically, the instruction-tuned LLM backbone provides a baseline tendency toward natural, paraphrased output over verbatim copying. This tendency is amplified by the decompose-then-refine design: partitioning full-text articles into facet-specific subdocuments allows each summarization step to operate over shorter, topically coherent contexts, reducing verbatim copying under long-context pressure. The subsequent refinement stage then resynthesizes the concatenated facet summaries, further encouraging paraphrasing over fragment assembly.

The distraction effect observed under DI and SI prompts indicates that instruction complexity can actively harm factual grounding in small LLMs. Similarly, entity guidance via UMLS only seemed to increase instruction complexity, yielding little positive effect. The inconsistent gains of SI+CoT further suggest that publication-type distribution can influence generation behavior. This motivates publication-type-aware prompting as a future direction. Additional BC-prompt ablations corroborate these findings for TR-UMLS (Appendix H).

To evaluate the generalizability of DPR-BAG, we applied it to the PubMed Summarization dataset (Cohan et al., 2018). Unlike PMC-MAD, this dataset is not stratified to reflect the publica-

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	0.605 (-0.058)	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
DI	0.725 (+0.063)	3.131 (-3.519)	0.642	0.806	**0.638**	0.885	0.247
DI+Top-5	0.725 (+0.062)	3.138 (-3.512)	0.644	0.807	0.637	0.933	0.245
DI+Top-10	**0.677 (+0.014)**	2.938 (-3.712)	0.644	0.811	0.595	0.915	0.231

Table 5: Ablation on UMLS entity guidance variants. Top-n denotes the number of top-ranked UMLS entities injected into the DI prompt. (Abbreviations as in Table 1 and Table 2)

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	0.828	**0.266**
SI	0.736 (+0.073)	3.049 (-3.601)	0.630	0.805	**0.636**	0.836	0.251
SI + CoT	0.749 (+0.086)	2.940 (-3.710)	0.596	0.792	0.631	**0.794**	0.249

Table 6: Ablation on Chain-of-Thought guidance for the SI prompt. (Abbreviations as in Table 1 and Table 2)

tion type distribution of abstract-less PubMed articles. Overall, the trends observed on PMC-MAD, particularly with respect to semantic alignment, abstractiveness, and factuality, largely persist on this dataset. Some differences are also observed (lower MiniCheck scores and better performance of SI+CoT relative to BC). Despite these variations, the results indicate that DPR-BAG maintains robust performance across datasets, supporting its generalizability (Appendix F).

To investigate whether larger model sizes can improve generation quality, we also evaluated `Qwen2.5:7B` and `Qwen2.5:14B` under the BC prompt. These models achieve improvements in semantic alignment and compression rates; however, they also yield excessive abstractiveness and lower factuality, overall yielding little advantage over the base 3B model (Appendix G).

6.1 Limitations

Several limitations apply. First, our evaluation relies primarily on automated metrics, which vary in their robustness. Human evaluation is needed for complementary validation. Second, the BOMRC schema may be suboptimal for articles with non-standard discourse structure. More adaptive decomposition strategies remain an open direction. Finally, DPR-BAG's summaries exhibit substantially higher compression than human-written abstracts, indicating considerably terser outputs. This excessive compression risks omitting auxiliary but informative content such as background, secondary findings, or caveats. Calibrating facet-level length targets is a natural direction for future work.

7 Related Work

For the BAG task, Chachra et al. (2016) integrated domain-specific classifiers and entailment graphs for extractive sentence selection, inheriting the disjointed flow typical of pure extractive approaches. Sybrandt and Safro (2021) proposed CBAG, which generates abstracts conditioned on author-provided MeSH keywords rather than full-text articles, leaving long-context challenges unaddressed.

To generate abstractive summaries from full-text articles, DANCER (Gidiotis and Tsoumakas, 2020) and IDCUOT (Shen and Lam, 2022) pioneered the strategy of breaking long documents into manageable sections to bypass context window limitations. However, these supervised approaches rely on heuristic alignment algorithms to map abstract sentences back to source sections and generate summaries for each section in isolation, making them prone to error propagation and fragmented coherence across sections.

GenCompareSum (Bishop et al., 2022) leverages similar divide-and conquer principles but remains extractive-heavy: abstractive fragments serve only as anchors for sentence selection, resulting in discontinuous summaries that lack the narrative transitions typical of human-written abstracts.

8 Conclusion

We presented DPR-BAG, a training-free, rhetorical structure-aware, divide-and-conquer framework for biomedical abstract generation. The method decomposes full-text articles into semantic facets and applies parallel LLM-based summarization followed by refinement. Across both the PMC-MAD and PubMed Summarization datasets, DPR-BAG produces abstracts that most closely match human-

written abstracts in terms of abstractiveness and factual consistency, without task-specific training and within standard hardware constraints. The framework can be integrated as a preprocessing component into pipelines that require biomedical abstracts, thereby improving downstream performance.

Acknowledgement

This work was supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM14292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. GenCompareSum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240, Dublin, Ireland. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Suchet Chachra, Asma Ben Abacha, Sonya Shooshan, Laritza Rodriguez, and Dina Demner-Fushman. 2016. A hybrid approach to generation of missing abstracts in biomedical literature. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1093–1100, Osaka, Japan. The COLING 2016 Organizing Committee.

Daniil Chernyshev and Boris Dobrov. 2024. Investigating the pre-training bias in low-resource abstractive summarization. *IEEE Access*, 12:47219–47230.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Brandon Fan, Weiguo Fan, Carly Smith, and Harold "Skip" Garner. 2020. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*, 57(1):102131.

Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, 13(13).

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on LLM performance? *Preprint*, arXiv:2411.10541.

John PA Ioannidis and Michaéla C Schippers. 2025. In-house editorials and journalistic pieces comprise a massive corpus in the scientific literature that can be improved. *European Journal of Clinical Investigation*, 55(8):e70061.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sylvey Lin, Joseph Menke, Arthur Holt, Halil Kilicoglu, and Neil Smalheiser. 2025. Section header normalization in biomedical articles using transformers. In *AMIA Annual Symposium Proceedings*. Poster P116.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022. Improving biomedical information retrieval with neural retrievers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11038–11046.

Anne Magnet and Didier Carnet. 2006. Letters to the editor: Still vigorous after all these years?: A presentation of the discursive and linguistic features of the genre. *English for Specific Purposes*, 25(2):173–199.

Joe D Menke, Halil Kilicoglu, and Neil R Smalheiser. 2024. Publication type tagging using transformer models and multi-label classification. *AMIA Annual Symposium Proceedings*, 2024:818–827.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

James L. Nuzzo. 2021. Letters to the editor in exercise science and physical therapy journals: an examination of content and "authorship inflation". *Scientometrics*, 126(8):6917–6936.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xin Shen and Wai Lam. 2022. Improved divide-and-conquer approach to abstractive summarization of scientific papers. In *2022 4th International Conference on Natural Language Processing (ICNLP)*, pages 395–398.

Justin Sybrandt and Ilya Safro. 2021. CBAG: Conditional biomedical abstract generation. *PLoS One*, 16(7):e0253905.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Alberto Ueda, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2021. Structured fine-tuning of contextual embeddings for effective biomedical retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2031–2035, New York, NY, USA. Association for Computing Machinery.

Cathelijn J. F. Waaijer, Cornelis A. van Bochove, and Nees Jan van Eck. 2011. On the map: Nature and science editorials. *Scientometrics*, 86(1):99–112.

Tairan Wang, Xiuying Chen, Qingqing Zhu, Taicheng Guo, Shen Gao, Zhiyong Lu, Xin Gao, and Xiangliang Zhang. 2025. New paradigm for evaluating scholar summaries: A facet-aware metric and a meta-evaluation benchmark. *ACM Trans. Inf. Syst.*, 43(4).

Thomas C Wiegers, Allan Peter Davis, Jolene Wiegers, Daniela Sciaky, Fern Barkalow, Brent Wyatt, Melissa Strong, Roy McMorran, Sakib Abrar, and Carolyn J Mattingly. 2025. Integrating AI-powered text mining from PubTator into the manual curation workflow at the Comparative Toxicogenomics Database. *Database*, 2025:baaf013.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yizhou Zhang, Defu Cao, Lun Du, Qiang Fu, and Yan Liu. 2025. When splitting makes stronger: A theoretical and empirical analysis of divide-and-conquer prompting in LLMs. In *Second Conference on Language Modeling*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

A Summarization Prompt Templates

All prompting strategies share a `facet_guidelines` dictionary that maps each facet label to a facet-specific instruction. We use two versions: a concise version (Table 7) used in BC and SI prompts, and a detailed version adapted from JMIR author guidelines (see footnote 1) (Table 8 used in DI prompts). In all prompt templates, `<facet_text>` denotes the input facet text, `<facet_type>` denotes the rhetorical facet label, and `<facet_guide>` denotes the corresponding facet-specific instruction.

A.1 Basic Concise Prompt (BC)

System Message

You are a biomedical summarization assistant. 1. Use a formal, objective, scientific tone. 2. Never use meta-phrases like 'the authors state' or 'this section describes'. Respond ONLY with a JSON object: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Summarize this `<facet_type>` section
Specific focus: `<facet_guide>`

Paragraph text:
`<facet_text>`

A.2 Detailed Instruction Prompt (DI)

System Message

You are a biomedical synthesis assistant. 1. Use a formal, objective, scientific tone.
2. Never use meta-phrases like 'the authors state' or 'this section describes'.
Define your output strictly as:
-'summary': The synthesized biomedical text.
-'reasoning': A brief explanation.
Format: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Synthesize the critical information from this `<facet_type>` section provided in the 'Paragraph text'
`<facet_guide>`

Paragraph text:
`<facet_text>`

A.3 Structural Instruction Prompt (SI)

System Message

ROLE
You are an expert Biomedical Summarization Assistant.

STRICT GUIDELINES
- **Tone**: Formal, objective, and academic.
- **No Meta-Talk**: Do NOT use phrases like 'The authors state' or 'This section describes'.
- **Output Format**: Respond **ONLY** with a valid JSON object. No Markdown blocks, no preamble, no postscript.
```json
{"reasoning": "Brief explanation of how your summary fulfills the given instructions...", "summary": "Final summary..."}
```

Facet	Guideline
Background	Focus on research gap and motivation.
Objective	State the primary aim or hypothesis.
Methods	Detail study design and procedures.
Results	Prioritize key findings and data.
Conclusions	Summarize implications and take-home messages.
Others	Focus on the main point of the paragraph.
Intro	Introduce the main topic and context.
Main Idea	Focus on the central concept or hypothesis.
Results & Conclusions	Focus on key findings and their implications.

Table 7: Concise facet-specific guidelines (used in BC and SI prompts).

Facet	Guideline
Background	Briefly describe the context and significance of the research.
Objective	State the specific aim(s) of the study in a complete sentence.
Methods	Outline the research design, study sample, data collection, and analysis procedures.
Results	Present key findings, including relevant statistics (sample sizes, response rates, P values, confidence intervals). Be specific.
Conclusions	Summarize the main findings and their implications.
Others	Synthesize the core biomedical information, focusing on the primary argument, concept, or supplementary context presented.
Intro	Describe the research context and significance, and clearly state the specific aims or hypotheses.
Main Idea	Outline the research design and procedures, while capturing any supplementary methodological context or core concepts.
Results & Conclusions	Present key findings with relevant statistics, and summarize their broader implications and take-home messages.

Table 8: Detailed facet-specific guidelines (used for DI prompts), adapted from JMIR author guidelines.

User Message

TASK: SUMMARY GENERATION
Target Focus: `<facet_guide>`

Instructions: Generate a professional biomedical summary.
—
INPUT TEXT (Reference)
`<facet_text>`

A.4 BC for Naive Splitting (BC-NS)

When paired with the Naive Splitting baseline, the system message remains unchanged. As no facet label is assigned, the user message uses a generic focus instruction:

User Message

Summarize this section.
Specific focus: Summarize the main topic.

Paragraph text:
`<facet_text>`

A.5 BC with TR-UMLS Entity Guidance (BC+TR-UMLS)

System Message

You are a biomedical summarization assistant. 1. Use a formal, objective, scientific tone. 2. Never use meta-phrases like 'the authors state' or 'this section describes'. Respond ONLY with a JSON object: {"summary": "...", "reasoning": "..."}. No markdown, no talk.

User Message

Summarize this `<facet_type>` section
Specific focus: `<facet_guide>`
Ensure the core meanings of these key biomedical entities are preserved or synthesized accurately:

Paragraph text:<top_entities>
`<facet_text>`

A.6 DI with TR-UMLS Entity Guidance (DI+TR-UMLS)

User Message

Synthesize the critical information from this <facet_type> section provided in the 'Paragraph text'
<facet_guide>

Ensure the core meanings of these key biomedical entities are preserved or synthesized accurately: <top_entities>

Paragraph text:
<facet_text>

A.7 SI with Chain-of-Thought (SI+CoT)

System Message

ROLE
You are an expert Biomedical Summarization Assistant.

—
OPERATIONAL FRAMEWORK
You must follow this 2-Stage Chain-of-Thought process:
1. Stage 1: Element Extraction (entities, parameters, methodologies, statistics).
2. Stage 2: Summary Generation (synthesize into scientific narrative).

—
STRICT GUIDELINES
- **Tone**: Formal, objective, and academic.
- **No Meta-Talk**: Do NOT use phrases like 'The authors state' or 'This section describes'.
- **Output Format**: Respond **ONLY** with a valid JSON object. No Markdown blocks, no preamble, no postscript.
```json
{"reasoning": "Stage 1 extraction results...", "summary": "Stage 2 final summary..."}
```

User Message 1 (Stage 1)

TASK 1: ELEMENT EXTRACTION
Section Type: <facet_type>
Instructions: Extract key elements from the text below, including:
- **Entities:** Diseases, genes, drugs, proteins.
- **Parameters:** Sample sizes, dosage, duration.
- **Methodology:** Study design, assays, equipment.
- **Statistics:** P-values, confidence intervals, effect sizes.

—
INPUT TEXT
<facet_text>

User Message 2 (Stage 2)

TASK 2: SUMMARY GENERATION
Target Focus: <facet_guide>

Instructions: Using the elements extracted in Task 1, generate a professional biomedical summary.
Ensure the summary is dense with information but remains readable and scientifically accurate.

—
INPUT TEXT (Reference)
<facet_text>

B Validation and Fallback Details

When the fallback process is triggered, the original six facets are regrouped into three broader categories: *Intro* (Background and Objective), *Main Idea* (Methods and Others), and *Results & Conclusions* (Results and Conclusions). Background and Objective are merged as both establish the research context and motivation. Results and Conclusions are paired as both convey findings and their implications. The remaining Others facet, which retains paragraphs not assigned to any BOMRC category by the sentence classifier, is grouped with Methods by elimination, as the other four facets form more natural rhetorical pairs. These regrouped facets are then sent back to the parallel summarization module. If the LLM still fails to summarize a specific

regrouped facet, the first 300 characters (Lead-300) of that facet are used as the facet summary to capture core information via lead bias without adding granular noise. The fallback mechanism was rarely triggered in practice, suggesting that this pairing has minimal impact on overall generation quality. In a random sample of 300 test articles (6.5% of the test set), the fallback mechanism was never triggered, suggesting with 95% confidence that fewer than 1% of articles require fallback intervention.

C Refinement Prompt

This prompt is used in the final stage to smooth the concatenated facets, ensuring structural coherence and stylistic consistency across the unified abstract. `<draft_abstract>` denotes the place holder for the concatenated draft abstract.

System Message

You are a biomedical abstract refinement assistant. Refine the abstract based on the abstract draft.
CRITICAL INSTRUCTION:
Respond ONLY with a valid JSON object.
Do NOT use Markdown code blocks (like “‘json).
Do NOT provide any conversational text.

Format:
{
"abstract": "your abstract text here",
"reasoning": "your reasoning here"
}

User Message

abstract draft: `<draft_abstract>`

D Extended Implementation Details and Token Distribution

All fine-tuning and evaluation procedures were conducted on an NVIDIA Tesla V100 GPU (32GB VRAM). While the underlying LED architecture theoretically supports sequences up to 16,384 tokens, processing such lengths on standard hardware is computationally prohibitive, inevitably leading to out-of-memory (OOM) errors even with minimal batch sizes. To fit within this memory budget, the fine-tuned LED-Pubmed used gradient accumulation with an effective batch size of 8, halted at

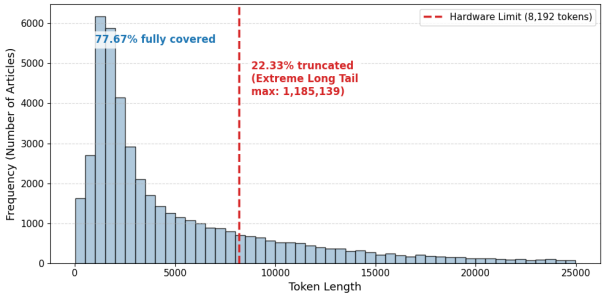


Figure 2: Distribution of document token lengths in the dataset. The red dashed line denotes the 8,192-token hardware limit for standard baselines.

500 steps based on validation performance.

This constraint restricts standard baselines to 8,192-token inputs. As illustrated in Figure 2, our empirical analysis of the 46,309 articles in the dataset reveals a median of 2,959 tokens and an average length of approximately 6,018 tokens. While the 8,192-token capacity successfully accommodates 77.67% of the dataset, the length distribution exhibits a severe long-tail characteristic. Specifically, 22.33% of the articles exceed this limit, with the longest document reaching an extreme 1,185,139 tokens. For these extensive studies, standard baselines operating within memory limits are forced to truncate critical information, such as discussion and conclusion sections, which underscores the necessity of the partitioned approach introduced in our DPR-BAG framework.

E LLM-SSC

LLM-SSC (Lan et al., 2024) is evaluated on the BIORC800 dataset, a manually annotated multi-label SSC dataset of biomedical abstracts using the BOMRC schema. Under task-specific fine-tuning, LLM-SSC achieves a micro F1 of 0.907 and macro F1 of 0.912 on BIORC800, outperforming prior SSC baselines (Cohan et al., 2019). We adopt LLM-SSC for our document splitting module as its label schema (Background, Objective, Methods, Results, Conclusions, and None) directly aligns with the BOMRC+ facets in DPR-BAG, where the None label corresponds to our Others facet.

F PubMedSum Dataset Validation

To evaluate the generalizability of our pipeline, we also tested DPR-BAG on the PubMed Summarization dataset (Cohan et al., 2018), hereafter PubMedSum. As shown in Figure 4, unlike PMC-MAD, which is stratified to reflect the publication type dis-

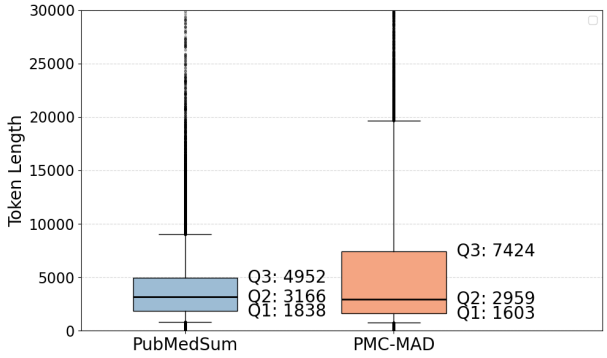


Figure 3: Token Distribution Comparison

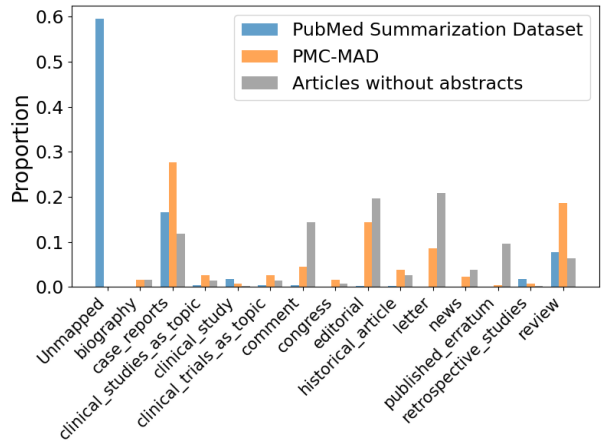


Figure 4: Publication type distribution of PMC-MAD, PubMed Summarization Dataset, and PubMed articles without abstracts.

tribution of abstract-less PubMed manuscripts, the majority of articles in the PubMedSum could not be mapped to a specific publication type. This allows us to benchmark the performance of DPR-BAG against established standards in the biomedical domain and ensure that our findings are not limited to the specific characteristics of the PMC-MAD corpus.

As illustrated in Figure 3, the two datasets exhibit distinct token length profiles. PMC-MAD has a median document length of 2959 tokens (IQR: 1603–7424), while PubMedSum has a median of 3,166 tokens (IQR: 1838–4952).

Table 9 and Table 10 present results on the PubMed Summarization dataset for the two best PMC-MAD configurations: BC (without entity guidance) and SI+CoT. Consistent with PMC-MAD findings, DPR-BAG underperforms all baseline models on ROUGE, UMLS Recall, DiscoScore, and SummaC, while achieving better coverage, density, and novel n-gram scores closer to the original human-written abstract. Both configurations outperform all baselines on AlignScore, confirming that the abstractiveness gains do not compromise factuality.

However, DPR-BAG's MiniCheck advantage on PMC-MAD does not transfer here, attributable to PubMedSum's inherent higher abstractiveness (lower density, higher novelty), which penalizes NLI-based metrics (MiniCheck) even when factual content is preserved — as evidenced by DPR-BAG's consistently superior AlignScore across both datasets.

Unlike on PMC-MAD where SI+CoT only improved BERTScore, on PubMedSum SI+CoT significantly improves AlignScore, ROUGE-L, UMLS Recall, DiscoScore, and Compression while showing no significant degradation on factuality (MiniCheck, SummaC) or abstractiveness. This suggests that CoT's benefit varies with document characteristics, potentially driven by publication-type distribution differences.

G Effect of Backbone Model Size

To investigate whether scaling up the backbone LLM improves generation quality, we evaluated `Qwen2.5:7B` and `Qwen2.5:14B` under the BC prompt without entity guidance. As the Llama3.2 series does not offer a model beyond 3B, we use Qwen 2.5 for the scaling analysis.

As shown in Table 11 and 12, Qwen2.5:14B improves BERTScore, UMLS Recall, and DS-Focus relative to the 3B backbone, recording the best DS-Focus, and UMLS Recall among all DPR configurations, alongside a compression rate most closely aligned with the original abstracts. However, both larger backbones exhibit excessively higher abstractiveness (lower Density, higher Novel n-grams) and declining factuality (SummaC, AlignScore and MiniCheck) relative to the 3B backbone, suggesting that larger models promote more abstractive phrasing and precise semantic alignment of key noun foci at the cost of reduced factual overlap with the source.

H BC Entity Guidance Ablations

To isolate the effect of entity guidance from prompt complexity, we additionally apply TR-UMLS to the BC prompt and compare against BC without entity guidance.

We integrate TR-UMLS with top-5 UMLS entities into the BC prompt (BC+Top-5). As shown in Table 13, BC+Top-5 significantly degrades fac-

Model	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.478	0.680	4.917	–	–	–
LED-Arxiv (base)	0.121 (-0.358)	0.205 (-0.475)	25.146 (+20.228)	0.755	**0.872**	**0.712**
LED-Pubmed (base)	0.156 (-0.322)	0.260 (-0.420)	19.565 (+14.648)	0.640	0.807	0.661
LongT5 (base)	0.157 (-0.322)	0.271 (-0.409)	15.827 (+10.910)	0.605	0.826	0.645
DPR-BAG (BC)	0.507 (+0.029)	0.733 (+0.053)	2.966 (-1.951)	0.765	0.794	0.414
DPR-BAG (SI+CoT)	**0.503 (+0.024)**	**0.723 (+0.043)**	**3.153 (-1.765)**	**0.772**	0.787	0.411

Table 9: Performance comparison on the PubMed Summarization dataset across primary evaluation dimensions. Conventions and abbreviations as in Table 1

Model	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.878	16.084
LED-Arxiv (base)	0.641	0.913	1.398	0.236	0.348	0.968 (+0.091)	**17.739 (+1.655)**
LED-Pubmed (base)	0.664	**0.929**	**1.100**	**0.272**	**0.415**	0.960 (+0.082)	14.235 (-1.849)
LongT5 (base)	**0.664**	0.894	1.585	0.265	0.345	0.962 (+0.085)	25.643 (+9.560)
DPR-BAG (BC)	0.651	0.814	2.116	0.192	0.217	0.876 (-0.002)	43.962 (+27.878)
DPR-BAG (SI+CoT)	0.652	0.839	1.927	0.197	0.245	**0.878 (+0.000)**	40.068 (+23.984)

Table 10: Performance comparison on the PubMed Summarization dataset across semantic alignment and supporting metrics. Conventions and abbreviations as in Table 2

tuality relative to BC ($p < 0.001$) and decreases UMLS Recall ($p < 0.001$), indicating that the injected UMLS terms fail to anchor concept reproduction in the output. These results confirm that TR-UMLS entity guidance remains ineffective in this zero-shot setting regardless of the underlying prompt strategy. Significance details are reported in Appendix J.5.

I Evaluation Metric Details

I.1 Factual Consistency (AlignScore, MiniCheck and SummaC)

AlignScore (Zha et al., 2023): AlignScore trains a unified alignment model on 4.7M examples spanning 7 tasks—NLI, fact verification, paraphrase, semantic textual similarity, question answering, information retrieval, and summarization—producing a single factual consistency score. At inference time, the source document is split into overlapping chunks of approximately 350 tokens, and each sentence in the generated abstract is evaluated against all chunks; the highest alignment score per sentence is averaged to yield the final score. We use the base variant (RoBERTa-base, 125M parameters).

MiniCheck (Tang et al., 2024): MiniCheck trains a small fact-checking model on synthetically constructed data generated by GPT-4, where each instance is designed to require verifying multiple atomic facts against multi-sentence evidence. It produces a binary supported/unsupported prediction per sentence. In our evaluation, each sentence in the generated abstract is treated as an individual claim and verified against the source full-text document. We use the Flan-T5-Large variant (770M parameters).

SummaC (Laban et al., 2022): SummaC segments the source document into individual sentences and scores each generated sentence by aggregating sentence-level NLI entailment probabilities against all source sentences. The $\text{SummaC}_{\text{Conv}}$ variant learns a convolutional layer over the full distribution of these entailment scores, rather than relying only on the maximum, making it more robust to outliers. We use the `SummaCConv` implementation with a VitaminC-trained NLI backbone and sentence-level granularity.

I.2 DiscoScore (DS_SENT_NN & DS_FOCUS_NN)

DiscoScore(Zhao et al., 2023) evaluates discourse coherence by modeling focus transitions across sentences. In this work, we use nouns (NN) as the focus, one of several focus choices supported by the metric, to compare the discourse coherence between the reference and generated abstracts.

DS_SENT_NN: Constructs a sentence graph where edges are drawn between any two sentences (not just adjacent ones) that share at least one noun focus. Edge weights are inversely proportional to the distance between the two sentences ($1/(j-i)$), capturing both local and long-range co-

Backbone	Abstractiveness			Factuality		
	Bi-g	Tri-g	Dens.	AS	MC	SummaC
Original abstract	0.495	0.663	6.650	–	–	–
Llama 3.2:3B	0.397 (-0.098)	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	**0.570**
Qwen 2.5:7B	0.594 (+0.099)	0.806 (+0.143)	2.380 (-4.270)	0.635	0.793	0.432
Qwen 2.5:14B	**0.568 (+0.073)**	0.778 (+0.115)	2.651 (-3.999)	0.647	0.812	0.443

Table 11: Scale-up comparison on primary evaluation dimensions. All configurations use the BC prompt without entity guidance. Conventions and abbreviations as in Table 1

Backbone	Semantic Alignment			Supporting			
	BS	SENT	FOCUS↓	R-L	U-R	Cov.	Comp.
Original abstract	–	–	–	–	–	0.743	23.827
Llama 3.2:3B	0.617	0.868	0.828	0.183	0.266	0.894 (+0.151)	50.037 (+26.210)
Qwen 2.5:7B	0.631	**0.876**	0.670	0.177	0.279	**0.846 (+0.103)**	32.273 (+8.446)
Qwen 2.5:14B	**0.633**	0.866	**0.596**	**0.189**	**0.337**	0.857 (+0.114)	**23.234 (-0.593)**

Table 12: Scale-up comparison on semantic alignment and supporting metrics. Conventions and abbreviations as in Table 2

herence. Sentence embeddings are then aggregated according to this graph structure, and the cosine similarity between the resulting graph-level embeddings of the generated and reference texts is used as the final score.

DS_FOCUS_NN: Measures how closely the frequency and semantics of shared noun foci match between the generated and reference texts. For each focus shared by both texts, it computes the distance between their embeddings (derived by summing the contextualized token embeddings of all associated tokens), and averages these distances as the final score.

I.3 UMLS Recall

This metric measures how well the generated abstract preserves biomedical concepts present in the reference abstract. We extract unique UMLS concepts from both the original abstract (C_{ref}) and the generated abstract (C_{gen}) using a biomedical entity linker (scispaCy, Neumann et al., 2019) with the UMLS knowledge base). UMLS Recall is then computed as:

$$\text{UMLS Recall} = \frac{|C_{\text{gen}} \cap C_{\text{ref}}|}{|C_{\text{ref}}|} \quad (1)$$

Newsroom (Grusky et al., 2018): Coverage, Density, and Compression: We adopt the extractive fragment measures from Grusky et al. (2018). Given a summary S and source document D, extractive fragments are the set of longest common substrings shared between S and D. **Coverage** measures the proportion of summary tokens belonging to an extractive fragment. **Density** measures the average length of the extractive fragments, reflecting how verbatim the summary is. **Compression** is the ratio of source document length to summary length.

N-gram Novelty Following See et al. (2017), we compute the proportion of bigrams and trigrams in the generated abstract that do not appear in the source full-text document. Higher novelty indicates greater abstractive synthesis relative to the source.

J Statistical Significance Analysis

J.1 Method

For each comparison between two configurations A and B, we align per-document scores by article ID and resample with replacement over 10,000 iterations. For metrics where higher (or lower) scores indicate better performance (AlignScore, MiniCheck, SummaC, BERTScore, DiscoScore variants, ROUGE-L, UMLS Recall), we test $H_0 : \mathbb{E}[s_A] = \mathbb{E}[s_B]$ on raw scores. For abstractiveness metrics where the goal is proximity to the reference distribution (Bigram novelty, Trigram novelty, Density, Coverage, Compression), we instead test $H_0 : \mathbb{E}[\,|s_A - s_{\text{ref}}|\,] = \mathbb{E}[\,|s_B - s_{\text{ref}}|\,]$, where s_{ref} is the corresponding score of the human-written abstract on the same document. Under this formulation, a negative difference indicates that A is closer to the reference than B. All tests are two-sided, with significance markers * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Comparisons not marked are not significant at $p < 0.05$. The per-document scores used for testing are extracted from the same evaluation pipeline that produces the average numbers in Section 5; only articles with

Config	Abstractiveness		Factuality		Semantic Alignment		
	Tri-g	Dens.	AS	MC	BS	FOCUS↓	U-R
BC	**0.605 (-0.058)**	**4.690 (-1.960)**	**0.762**	**0.890**	0.617	**0.828**	**0.266**
BC+Top-5	0.733 (+0.070)	2.994 (-3.656)	0.637	0.798	0.639	0.931	0.238

Table 13: Ablation of TR-UMLS entity guidance applied to the BC prompt. (Abbreviations as in Table 1 and Table 2)

valid scores from both configurations are included in each pairwise test.

In all tables below, mean differences are computed as $A - B$. Arrows next to metric names indicate whether higher (↑) or lower (↓) values are preferred.

J.2 DPR-BAG vs. Baselines (PMC-MAD)

Table 14, Table 15, and Table 16 report significance for the comparisons between DPR-BAG (BC) and the baseline models on PMC-MAD.

Metric	vs. LED-arXiv	vs. LED-PubMed
Bi-g↓	-0.152^{***}	-0.131^{***}
Tri-g↓	-0.217^{***}	-0.187^{***}
Dens.↓	-12.121^{***}	-9.048^{***}
AS↑	$+0.041^{***}$	$+0.095^{***}$
MC↑	$+0.033^{***}$	$+0.044^{***}$
SummaC↑	-0.167^{***}	-0.124^{***}
BS↑	-0.019^{***}	-0.036^{***}
SENT↑	-0.039^{***}	-0.040^{***}
FOCUS↓	$+0.126^{***}$	$+0.158^{***}$
R-L↑	-0.052^{***}	-0.080^{***}
U-R↑	-0.075^{***}	-0.089^{***}
Cov.↓	-0.017^{***}	-0.011^{***}
Comp.↓	$+20.499^{***}$	$+20.720^{***}$

Table 14: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the pretrained LED baselines on PMC-MAD.

Metric	vs. LED-PubMed-FT	vs. LongT5
Bi-g↓	-0.032^{***}	-0.128^{***}
Tri-g↓	-0.042^{***}	-0.170^{***}
Dens.↓	-2.383^{***}	-5.565^{***}
AS↑	$+0.251^{***}$	$+0.093^{***}$
MC↑	$+0.236^{***}$	$+0.013^{***}$
SummaC↑	$+0.067^{***}$	-0.142^{***}
BS↑	-0.018^{***}	-0.048^{***}
SENT↑	-0.000	-0.008^{***}
FOCUS↓	$+0.033$	-0.036
R-L↑	-0.082^{***}	-0.093^{***}
U-R↑	-0.099^{***}	-0.047^{***}
Cov.↓	$+0.052^{***}$	-0.015^{***}
Comp.↓	$+22.307^{***}$	$+6.213^{*}$

Table 15: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the fine-tuned LED-PubMed and LongT5 baselines.

Metric	vs. LongT5-FT
Bi-g↓	-0.059^{***}
Tri-g↓	-0.081^{***}
Dens.↓	-3.048^{***}
AS↑	$+0.221^{***}$
MC↑	$+0.164^{***}$
SummaC↑	$+0.005$
BS↑	-0.034^{***}
SENT↑	-0.027^{***}
FOCUS↓	$+0.111^{***}$
R-L↑	-0.095^{***}
U-R↑	-0.097^{***}
Cov.↓	$+0.021^{***}$
Comp.↓	$+19.467^{***}$

Table 16: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and the fine-tuned LongT5 baseline.

J.3 Splitting Strategies

Table 17 reports significance for the comparisons between FS and other splitting strategies.

Metric	vs. NS	vs. SH
Bi-g↓	$+0.005^{*}$	$+0.045^{***}$
Tri-g↓	$+0.006^{*}$	$+0.046^{***}$
Dens.↓	$+0.150^{**}$	$+0.698^{***}$
AS↑	$+0.006^{*}$	$+0.126^{***}$
MC↑	-0.003	$+0.099^{***}$
SummaC↑	$+0.003$	$+0.141^{***}$
BS↑	$+0.001$	-0.019^{***}
SENT↑	$+0.004^{*}$	$+0.015^{***}$
FOCUS↓	-0.033	-0.098^{***}
R-L↑	$+0.002^{*}$	$+0.004^{***}$
U-R↑	$+0.004$	$+0.037^{***}$
Cov.↓	$+0.000$	$+0.057^{***}$
Comp.↓	$+1.875$	$+1.897$

Table 17: Mean differences between FS and other splitting strategies.

J.4 Prompting Strategies

Table 18 reports significance for the prompting strategy ablation.

J.5 TR-UMLS Entity Guidance

Table 19 and Table 20 report significance for the TR-UMLS entity guidance ablation.

Metric	BC vs. DI	BC vs. SI	DI vs. SI
Bi-g↓	$+0.041^{***}$	$+0.037^{***}$	-0.004^{**}
Tri-g↓	$+0.045^{***}$	$+0.040^{***}$	-0.005^{***}
Dens.↓	$+0.662^{***}$	$+0.616^{***}$	-0.045^{*}
AS↑	$+0.121^{***}$	$+0.132^{***}$	$+0.012^{***}$
MC↑	$+0.084^{***}$	$+0.086^{***}$	$+0.002$
SummaC↑	$+0.133^{***}$	$+0.139^{***}$	$+0.006^{**}$
BS↑	-0.022^{***}	-0.019^{***}	$+0.003^{***}$
SENT↑	$+0.007^{***}$	$+0.004^{**}$	-0.003
FOCUS↓	-0.065^{**}	-0.013	$+0.050^{**}$
R-L↑	$+0.001$	$+0.001$	$+0.001$
U-R↑	$+0.018^{***}$	$+0.014^{***}$	-0.004^{*}
Cov.↓	$+0.052^{***}$	$+0.061^{***}$	$+0.009^{***}$
Comp.↓	$+3.895$	$+5.627$	$+1.732^{*}$

Table 18: Mean differences across prompting strategies.

Metric	vs. DI+Top-5	vs. DI+Top-10
Bi-g↓	$+0.000$	-0.021^{***}
Tri-g↓	-0.000	-0.030^{***}
Dens.↓	-0.011	-0.137^{***}
AS↑	-0.003	-0.002
MC↑	-0.001	-0.006
SummaC↑	$+0.000$	-0.002
BS↑	$+0.001$	$+0.043^{***}$
SENT↑	$+0.002$	$+0.001$
FOCUS↓	-0.048^{**}	-0.026
R-L↑	-0.001	$+0.011^{***}$
U-R↑	$+0.002$	$+0.016^{***}$
Cov.↓	-0.002^{**}	-0.043^{***}
Comp.↓	$+0.395$	-0.069

Table 19: Mean differences between DI and TR-UMLS-augmented DI.

Metric	vs. BC+Top-5	vs. DI+Top-5	vs. DI+Top-10
Bi-g↓	$+0.040^{***}$	$+0.041^{***}$	$+0.021^{***}$
Tri-g↓	$+0.044^{***}$	$+0.045^{***}$	$+0.016^{***}$
Dens.↓	$+0.685^{***}$	$+0.650^{***}$	$+0.525^{***}$
AS↑	$+0.126^{***}$	$+0.118^{***}$	$+0.119^{***}$
MC↑	$+0.092^{***}$	$+0.083^{***}$	$+0.078^{***}$
SummaC↑	$+0.134^{***}$	$+0.133^{***}$	$+0.130^{***}$
BS↑	-0.022^{***}	-0.021^{***}	$+0.022^{***}$
SENT↑	$+0.014^{***}$	$+0.008^{***}$	$+0.008^{***}$
FOCUS↓	-0.107^{***}	-0.113^{***}	-0.087^{***}
R-L↑	$+0.002$	$+0.000$	$+0.012^{***}$
U-R↑	$+0.028^{***}$	$+0.021^{***}$	$+0.035^{***}$
Cov.↓	$+0.051^{***}$	$+0.050^{***}$	$+0.009^{***}$
Comp.↓	$+3.045$	$+4.291$	$+3.827$

Table 20: Mean differences between BC and TR-UMLS-augmented variants.

Metric	BC vs. SI+CoT	SI vs. SI+CoT
Bi-g↓	$+0.032^{***}$	-0.005^{**}
Tri-g↓	$+0.037^{***}$	-0.003
Dens.↓	$+0.591^{***}$	-0.030
AS↑	$+0.166^{***}$	$+0.034^{***}$
MC↑	$+0.098^{***}$	$+0.012^{**}$
SummaC↑	$+0.158^{***}$	$+0.019^{***}$
BS↑	-0.015^{***}	$+0.005^{***}$
SENT↑	-0.000	-0.004^{***}
FOCUS↓	$+0.031$	$+0.042^{*}$
R-L↑	$+0.005^{***}$	$+0.004^{***}$
U-R↑	$+0.017^{***}$	$+0.003$
Cov.↓	$+0.070^{***}$	$+0.009^{***}$
Comp.↓	$+7.488^{*}$	$+1.860^{***}$

Table 21: Mean differences for the SI+CoT ablation.

J.6 Chain-of-Thought Guidance

Table 21 reports significance for the SI+CoT ablation.

J.7 PubMedSum Generalization

Table 22 reports significance for DPR-BAG (BC) against the baseline models on PubMedSum, and Table 23 compares BC against SI+CoT on the same dataset.

J.8 Backbone Scale

Table 24 reports significance for the backbone scaling experiments.

K Case Study: Qualitative Comparison of Generated Abstracts

As shown in Table 25, both LED and LongT5 models exhibit severe verbatim copying from the source full text, with LongT5 further producing a lexical hallucination. In contrast, DPR-BAG demonstrates stronger topic identification and more abstractive generation for the BAG task, though it still occasionally exhibits entity relation confusion (e.g., misattributing target genes as lncRNAs).

Metric	vs. LED-arXiv	vs. LED-PubMed	vs. LongT5
Bi-g↓	-0.226^{***}	-0.190^{***}	-0.190^{***}
Tri-g↓	-0.333^{***}	-0.279^{***}	-0.269^{***}
Dens.↓	-17.806^{***}	-12.261^{***}	-8.710^{***}
AS↑	$+0.010^{**}$	$+0.125^{***}$	$+0.160^{***}$
MC↑	-0.078^{***}	-0.013^{**}	-0.032^{***}
SummaC↑	-0.299^{***}	-0.248^{***}	-0.231^{***}
BS↑	$+0.011^{***}$	-0.012^{***}	-0.013^{***}
SENT↑	-0.099^{***}	-0.115^{***}	-0.080^{***}
FOCUS↓	$+0.719^{***}$	$+1.015^{***}$	$+0.532^{***}$
R-L↑	-0.044^{***}	-0.081^{***}	-0.073^{***}
U-R↑	-0.131^{***}	-0.198^{***}	-0.127^{***}
Cov.↓	-0.035^{***}	-0.024^{***}	-0.027^{***}
Comp.↓	$+21.934^{***}$	$+23.572^{***}$	$+16.835^{***}$

Table 22: Mean differences (DPR-BAG minus baseline) between DPR-BAG (BC) and baseline models on PubMedSum.

Metric	BC vs. SI+CoT
Bi-g↓	-0.002
Tri-g↓	-0.001
Dens.↓	-0.016
AS↑	-0.007^{*}
MC↑	$+0.007$
SummaC↑	$+0.003$
BS↑	-0.000
SENT↑	-0.025^{***}
FOCUS↓	$+0.189^{***}$
R-L↑	-0.006^{***}
U-R↑	-0.028^{***}
Cov.↓	$+0.001$
Comp.↓	$+2.999^{***}$

Table 23: Mean differences between DPR-BAG (BC) and DPR-BAG (SI+CoT) on PubMedSum.

Metric	vs. Qwen-7B	vs. Qwen-14B
Bi-g↓	$+0.006$	$+0.043^{***}$
Tri-g↓	-0.005	$+0.046^{***}$
Dens.↓	$+0.519^{***}$	$+0.736^{***}$
AS↑	$+0.131^{***}$	$+0.115^{***}$
MC↑	$+0.098^{***}$	$+0.078^{***}$
SummaC↑	$+0.139^{***}$	$+0.128^{***}$
BS↑	$+0.048^{***}$	-0.015^{***}
SENT↑	-0.008^{***}	$+0.003$
FOCUS↓	$+0.145^{***}$	$+0.242^{***}$
R-L↑	$+0.023^{***}$	-0.006^{***}
U-R↑	$+0.013^{***}$	-0.070^{***}
Cov.↓	$+0.012^{***}$	$+0.068^{***}$
Comp.↓	$+15.960^{***}$	$+21.741^{***}$

Table 24: Mean differences (Llama-3.2:3B minus Qwen) between Llama-3.2:3B and the scaled Qwen2.5 backbones (denoted Qwen-7B and Qwen-14B).

Model	Generated Abstract (truncated)
Original Abstract	Long non-coding RNAs (lncRNAs) comprise a sizeable class of non-coding RNAs with a length of over 200 base pairs. Little is known about their biological function, although over 20,000 lncRNAs have been annotated in the human genome. Through a diverse range of mechanisms, their primary function is in the regulation of the transcription of protein-coding genes. lncrna transcriptional activation can result from a group of nucleus-retained and chromatin-associated lncrnas, which function as scaffolds in the cis/trans recruitment of transcription factors, co-activators or chromatin remodelers, and/or promoter enhancers. their synthesis is initiated by various processes including ceramide synthesis, release of intracellular ca2+ or acid-base balance disorders. prior to vesicle creation, selective cargo loading occurs in the endosomal sorting complex required for transport. participation of endosomal sorting proteins such as tetraspanins or specific sumoylated proteins required for transport has been indicated in research. the endosomal-sorting complex consists of four components, these induce the formation of multivesicular bodies and the induction of membrane deformation to form exosomes. nanovesicles could be formed inside multivesicular bodies to allow transport outside the cell or digestion in lysosomes. the molecular content of exosomes is more heterogenic than its synthesis process, with different cargoes being examined inside vesicles with regard to the type or stage of cancers. this paper will review the importance of lncrnas as crucial molecular content of exosomes, indicating its involvement in tumour suppression, pro-tumorigenic events and the development of novel therapeutic approaches in the near future. further studies of their mechanisms of function are essential, as well as overcoming several challenges to gain a clearer insight to the approaches for the best clinical application.
LED-Pubmed (FT)	The exosome is a type of extracellular vesicle involved in cell-to-cell communication and disease transmission. Exosomes can be found in urine, sperm, amniotic fluid and other biological fluids, and have been implicated in various biological processes, such as cancer development, immune responses, invasion, invasion and invasion, cell proliferation, invasion or invasion, migration and invasion of tumour cells, as well as in the regulation of gene expression in cancer cells. the lncrna (non-coding rna) is a class of transcripts that do not encode proteins and whose length exceeds 200 bp and therefore play a vital role in regulating signaling pathways and gene expression, and therefore plays a pivotal role in cancer development and cancer-related diseases such as metastasis, angiogenesis and tumour cell invasion, and is involved in the formation of exosomal lipid bilayers, migration, proliferation and invasion in tumour microenvironment, which in turn are involved in cancer progression and progression, as evidenced by their involvement in the secretion of cytokines and growth factors, lipids or proteins from endosomes, and in the activation of the transcription factor hypoxia inducible factor (hif-β) pathway, among others, and their role in the induction of pre-metastatic niche formation in cancerous diseases, including lung cancer. the purpose of this review is to summarize the role of various exosomally-derived lnas in carcinogenesis and in other biological processes and to highlight the importance of these molecules in cancer biology and cancer development.
LongT5 (Base)	Circulating long non-coding rnas are a class of transcripts that do not encode proteins and whose length exceeds 200 bp. They play a significant role in diverse cellular processes from enabling normal development to disease progression, as well as cell communication. A valuable example of a biological carrier that can transport abundant quantities of long non - codons is that of the exosome. its role has been documented in both physiological conditions and pathological changes such as cancer development, where they are responsible for the regulation of various processes including immunosuppression, proliferation and induction of pre - metastatic niche formation. in this review, the role of exosomal long non
DPR-BAG (BC)	LncRNA has greater tissue specificity than mRNA, making it a potential diagnostic biomarker for specific tumour types. Circulating RNA in body fluids can resist ribonucleases, providing stability and potential as cancer prognostic markers. A study found two groups of long non-coding RNAs (lncRNAs) with opposite effects on cancer cells. Gas5 was shown to induce apoptosis and inhibit tumour growth, while Particl enhanced DNA methylation and gene silencing. Another lncRNA, CCND1, bound to a modulator and blocked transcription factors. Exo1-4 lncRNAs increased recipient cell viability by interacting with proteins involved in metabolism and nucleosomal architecture. Exosomal long non-coding RNAs(lncRNAs) have shown potential as effective nanocarriers for regulating lncRNA expression in cancer cells, offering a promising tumour therapy approach.

Table 25: Qualitative comparison of generated abstracts for a sample article (PMCID: PMC6625196) from the PMC-MAD test set. Orange: verbatim copy from source full text. Red: factual error.

AAbAAC: An Annotated Corpus for Autoimmunity Information Extraction

Fabien Maury[1,2], Solène Grosdidier[3], Maud de Dieuleveult[1,*], Adrien Coulet[2,*]

[1]Inserm, Université Paris Cité, U1163 Institut Imagine, Paris, France,
[2]Inria, Inserm, Université Paris Cité, U1346 HeKA, Paris, France,
[3]Freelance researcher, The Hague, Netherlands,

[*]These authors contributed equally

Correspondence: fabien.maury@inserm.fr

Abstract

Despite advances in information extraction driven by deep learning and large language models, performance gaps remain in highly specialized biomedical fields, where domain-specific complexity poses challenges for generalist models. In this work, we focus on the domain of autoimmunity, where the main entities of interest are autoimmune diseases, autoantibodies (*i.e.*, molecules that may mark or cause these diseases), their molecular targets, their location in the body, and their associated clinical signs. Herein, we present AAbAAC (AutoAntibodies and Autoimmunity Annotated Corpus), a corpus of 115 abstracts selected from PubMed, where we manually annotated entities and their relationships. First, AAbAAC was used to evaluate several methods on the task of named entity recognition (NER), and secondly, to fine-tune NER models. Our study demonstrates the utility of AAbAAC for information extraction in the domain of autoimmunity, showing expected improvement in NER performance after fine-tuning. This illustrates the value of small-scale annotation efforts for specialized domains and contributes to the computational study of autoimmunity. The AAbAAC corpus is available at https://github.com/f-maury/AAbAAC.

1 Introduction

Autoimmune diseases arise from dysfunction of the adaptive immune system, whereby molecules of the self are targeted by self-antibodies, named autoantibodies. Those are generally detected in blood or cerebrospinal fluid, and widely used as a biomarker of autoimmune diseases, critically guiding differential diagnosis. Autoimmune diseases collectively affect an estimated 3–5% of the global population (Ahsan, 2023; Shapira et al., 2010). Yet, when considered individually, many of these diseases are rare (Hayter and Cook, 2012). In part due to the rarity of these pathologies, there is no single centralized resource including all relevant information about autoimmunity. Indeed, relevant bits of domain expertise are scattered across various sources, including many recent research papers. For this reason, information extraction is a relevant way to study autoimmune diseases, by automating the processing of large numbers of documents into structured, interoperable, and machine-readable knowledge bases.

Lately, BERT-based models have consistently achieved state-of-the-art performance in information extraction tasks, particularly when adapted through fine-tuning on domain-specific documents matching the target domain. While many domain-specific models exist, high-quality annotated data to adapt general models to highly specialized or niche domains are also lacking in many cases. To our knowledge, this is the case in human autoimmunity, where no manually annotated corpus dedicated to autoimmune diseases and autoantibodies exists to evaluate or fine-tune models specialized for the recognition of autoantibodies, autoimmune diseases, and their relationships. To fill this gap, we propose AAbAAC, a corpus of PubMed abstracts annotated for entities and relations related to autoimmunity, and demonstrate its value for named entity recognition (NER).

Section 2 reviews related works. Section 3 describes the methodology used to create the corpus, and Section 4 presents the resulting corpus. Section 5 introduces the experimental setup for named entity recognition (NER) using AAbAAC, and Section 6 reports the associated results. Finally, the paper concludes with perspectives for future work enabled by this new corpus.

2 Related works

2.1 Existing resources for autoimmunity

Some general and widely used knowledge bases and resources exist for broad biomedical

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 791–800
July 3-4, 2026 ©2026 Association for Computational Linguistics

use, such as the Human Phenotype Ontology (HPO) (Gargano et al., 2024), Orphanet (Rath et al., 2012), or the UMLS (Bodenreider, 2004). Those include information about autoimmune diseases and autoantibodies, but are not sufficiently exhaustive, detailed, nor up-to-date for real-world studies. Indeed, autoimmunity is a specialized and evolving topic that is not the focus of generalist resources. They can nonetheless be used to generate an initial list of autoantibody names, that can then be used for information extraction.

More specialized resources have also been developed that include information about antibodies, autoantibodies, or related elements. For example, this is the case of IEDB (Vita et al., 2025), a database about epitopes, *i.e.*, regions of antigens bound by antibodies, or AAgAtlas (Wang et al., 2017), which focuses on autoantigens. Each of these resources focuses on a topic that is relevant to the study of autoimmunity, but none of them compile all the relevant knowledge in the field of autoimmunity for clinicians or researchers to use. However, for information extraction purposes, these resources could be used to build lists of autoantibody names from their antigen names.

Some annotated text corpora already exist for biomedical information extraction, but to the best of our knowledge, none is dedicated to human autoimmunity and autoimmune diseases. For example, the MedMentions corpus (Mohan and Li, 2019) tackles the biomedical domain in a general manner, with annotations of UMLS terms of most Semantic Types. As a result, it may include the most common autoimmune diseases and autoantibodies, but it does not include uncommon ones that do not correspond to UMLS terms, nor does it include relationships between the entities. The ABAG corpus (Dinh et al., 2022) focuses on antibodies and antigens, but is not specific to autoimmunity and does not provide annotations of relationships between entities. Therefore, existing resources are of limited use for autoimmunity-related information extraction, since isolating in these resources the annotations specific to autoimmunity would be challenging and result in limited coverage of autoimmunity-related concepts.

2.2 NER for autoimmunity

Many works make use of dictionaries and rule-based methods to conduct NER with decent performance. Such dictionaries can be built from either ontologies or databases. For example, a recent work by Remaki et al. (2025) used such a combination, including terms from SNOMED CT (Wang et al., 2002) and rules, to extract biology exams and drugs in relations to immune-mediated inflammatory diseases.

Another work by Subramanian and Ganapathiraju (2017) extracts antibody names using simple regular expressions based on the fact that antibodies are often written as "X antibody", or "antibody to X", where "X" is the target of the antibody. This method does not require a dictionary of possible target names, but fails at covering the various possible ways an antibody can be mentioned in text. Moreover, in the context of autoimmunity, such a method would struggle to differentiate autoantibodies from regular antibodies.

The development of multi-head attention and transformer-based architectures of language models enables efficient information extraction from long texts with more flexibility and context awareness than rule-based approaches. For NER tasks, encoder models such as BERT-based models (Devlin et al., 2019) perform competitively while being lighter and less computationally expensive to work with than large language models (LLM) (Naguib et al., 2024). The original BERT model has been adapted to many domains, including the biomedical domain. However, we did not find any model specialized for the recognition of autoimmunity-related entity types.

3 Corpus creation methods

3.1 Text selection

To select a set of relevant texts to be annotated, we first built a dictionary of autoantibody names and synonyms from HPO terms, which we used to query a PubMed API (Sayers, 2018) for titles and abstracts of scientific papers. Indeed, HPO includes terms describing positivity to some autoantibody detection tests (terms under *HP:0030057*). Most of them are of the form: "*anti-X antibody positivity*", where "*anti-X antibody*" is the name of an autoantibody that targets the antigen "*X*". We relied on this regular syntax to extract names of autoantibodies and manually reviewed results. We also generated lexical variants by reducing each autoantibody name to a core string through the removal of selected prefixes and suffixes, then recombining this core with alternative affixes. For example, "*anti-smooth muscle antibody*" was reduced to "*smooth muscle*" by removing "*anti*" and "*antibody*", and

one resulting variant was "*smooth muscle autoantibody*". Overall, the dictionary includes a total of 10,916 variations for 285 unique autoantibodies that originated from HPO release 2024-07-01. On average, each autoantibody in the dictionary has 38.30 variations, including synonyms that were already listed in the HPO.

For each autoantibody type in the dictionary, PubMed is searched with a query of the form "*(""autoantibody_name_1""[Title/Abstract] OR ""autoantibody_name_2""[Title/Abstract] AND humans[MeSH Terms] AND (antibodies[MeSH Terms] OR autoantibodies[MeSH Terms])*". MeSH term filters were added to select only papers tagged with "human" and either "antibodies" or autoantibodies. This process returned a total of 56,750 titles and abstracts. The full query we used is available in the project's GitHub repository.

Despite MeSH terms filtering, some of the obtained texts were irrelevant due to not being written in English or not being related to autoimmunity. For this reason, we enforced additional filters: abstracts were automatically pre-annotated with exact matches of autoantibody names from our HPO-derived dictionary, and with disease and autoantibody names identified by GLiNER. This pre-annotation was performed with a version of the GLiNER specialized for the biomedical field ("*urchade/gliner_large_bio-v0.1*"). This model is an early version of GLiNER large that is finetuned on PubMed abstracts.

Only texts containing at least one HPO autoantibody term match, or one disease pre-annotation and one autoantibody pre-annotation were kept, reducing the set of texts to 44,890.

Among these, 120 texts were randomly selected for manual annotation, out of which 5 were manually eliminated due to either not being related to autoantibodies or not being English texts, resulting in a final set of 115 texts. The initial selection size was 120 because we aimed to have a corpus of around 100 texts and expected a few of them to be eliminated.

3.2 Annotation guidelines and process

Annotation was performed by 4 annotators with a background in either medical informatics, biology, or pharmacy. This group elaborated annotation guidelines through two preliminary annotation batches to identify main entity types, difficulties, and edge cases before the start of the annotation campaign (Hovy and Lavid, 2010). Annotation guidelines are available in the following GitHub repository: https://anonymous.4open.science/r/aabaac-FA3F. Five types of entities were annotated: "*Autoantibody*", "*Autoantibody target*", "*Disease*", "*Symptom or clinical sign*", and "*Autoantibody location*"; and ten different types of relationships: "*is associated with*", "*is not associated with*", "*may be associated with*", "*is caused by*", "*is target of*", "*is reference to*", "*discontiguous entity*" (this relation is used to link several separated parts of a single entity), "*is synonym of*", "*is located in*", and "*is subclass of*".

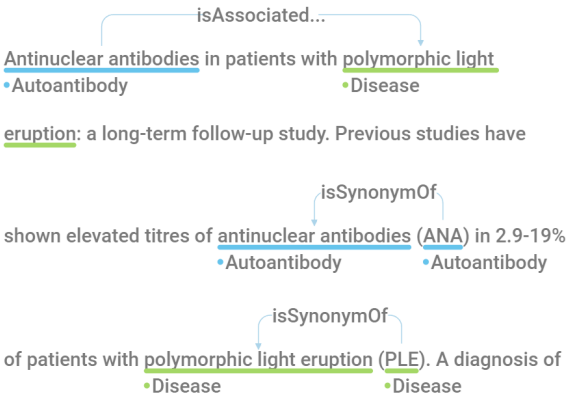


Figure 1: Example of text annotated in the Doccano web interface.

To make the annotation task easier for annotators, it was agreed to not annotate obvious relationships between nested entities, which were automatically added after manual annotation. For instance, when a "*Target*" is found inside the span of an "*Autoantibody*", which typically happens in structures like "*anti-X antibody*", the annotators were instructed not to add an "*is target of*" relation. Similarly, when an "*Autoantibody*" is found inside the span of a related "*Symptom or clinical sign*", the annotators were instructed not to add an "*is associated with*" relationship. Those were automatically added after manual annotation.

Each text was annotated independently by 2 annotators, but no single annotator annotated all texts. Each annotator was paired with each other annotator with equal frequency, and each pair annotated the same number of texts. Annotation was performed using Doccano (Nakayama et al., 2018) on texts where entities had been pre-annotated, but not relations. These pre-annotations were obtained by string matching and GLiNER. Doccano web interface is illustrated in Figure 1.

Texts were sent to the annotators in small numbered batches of around 4 texts, and annotators were free to annotate them in any order, in a single or multiple sessions, and at any time before the end of the campaign. Finally, an adjudication took place where annotations from both annotators for each text were confronted, and a consensual final annotation was decided by the head annotator.

4 Resulting corpus description

This section includes descriptive counts and statistics from the final adjudicated corpus of 115 texts. The variability of the length of the texts is high (from 50 to 3600 characters, as per Table 1), as is their annotation content (from 2 to 88 entities, and from 0 to 55 relations). This is partially due to the fact that some of the texts are limited to an article title with an empty abstract.

Metric	Min	Max	Average	Median
Entities per text	2	88	29.18	27
Relations per text	0	55	15.47	14
Text length (chars)	50	3600	1500.94	1500
Sentences per text	1	64	13.46	12

Table 1: General corpus statistics.

Imbalance is also observed in the number of entities of each type, with especially only 64 (see Table 2) occurrences of "*Autoantibody location*", whereas *Disease* is the most common entity with 1159 occurrences.

Entity type	Count	Avg/text	Median/text
Autoantibody	891	7.75	6
Autoantibody location	64	0.56	0
Autoantibody target	581	5.05	3
Disease	1159	10.08	8
Symptom or clinical sign	661	5.75	5
Total	3356	29.18	27

Table 2: Counts and distributions in the corpus by entity type.

Extremely variable numbers of occurrences can be observed for each relation type. Most relation annotations consist of "*is associated with*" (735, see Table 3), which can be explained by its rather broad definition, "*is target of*" (494), and "*is synonym of*" (226). Some relation types have almost never been used, such as "*is caused by*" (13), which could be explained by the cautious tone used in scientific literature, and the difficulty to establish causation with certainty, and "*is reference to*" (2), which could be explained by its constrained definition in our guidelines and redundancy with "*is synonym of*".

Relation type	Count	Avg/text	Median/text
discontiguousEntity	106	0.92	0
isAssociatedWith	735	6.39	5
isCausedBy	13	0.11	0
isLocatedIn	73	0.64	0
isNotAssociatedWith	24	0.21	0
isRefTo	2	0.02	0
isSubClassOf	61	0.53	0
isSynonymOf	226	1.97	1
isTargetOf	494	4.30	3
mayBeAssociatedWith	45	0.39	0
Total	1779	15.47	14

Table 3: Counts and distributions in the corpus by relationship type.

To evaluate the difficulty of the annotation task, inter-annotator agreement measures were computed between each pair of annotators, and between each annotator and the final annotations. Since the annotators simultaneously had to decide whether a text contained annotations, where in the text were these annotations located, if any, and of what type they were, Mathet's gamma coefficient (Mathet et al., 2015; Titeux and Riad, 2021) was used. Mathet's gamma is a measure that takes into account both disagreement on span limits (unitizing) and on label type (categorizing) for entity annotations. To measure agreement on relation annotations, three different comparison functions (strict match, flexible with tolerance for some disagreement, and super-flexible with tolerance for more disagreement) were used to compute F1-scores between the annotations of each annotator pair. We report a general average gamma coefficient between pairs of annotators of 0.67. The average gamma was 0.74 for "*Autoantibody*", 0.71 for "*Autoantibody location*", 0.44 for "*Autoantibody target*", 0.81 for "*Disease*", and 0.54 for "*Symptom or clinical sign*".

Relationship annotations are intrinsically more complex, and this is particulary the case in this work where annotators had 10 different relation types to choose from. To assess this difficulty quantitatively, we measured the inter-annotator agreement for relationships using an F1-score. We obtain a F1-score of 0.37 for all relation types together, using a custom flexible comparison metric tolerating some small overlap differences, and some entity type confusions.

Considering all relation types jointly, the primary source of disagreement was *isAssociatedWith*, because it was the most frequently used relation

type and because its broad definition allowed for subjective interpretation. When examining relation types individually, some exhibited lower agreement scores than *isAssociatedWith*; however, their contribution to the overall disagreement remained limited due to their lower frequency. In particular, *isRefTo* was used only twice and obtained an F1 score of 0. Given its limited use, removing this relation type from the annotation scheme could be considered. Overall, the relatively low inter-annotator agreement observed for relations highlights the difficulty of annotating texts from a specialized scientific domain. Furthermore, the complexity of the annotation scheme made the annotation task challenging for annotators.

5 Evaluating NER methods

5.1 Experimental setup

To demonstrate the utility of the AAbAAC corpus in autoimmunity NER, several methods were compared.

The corpus was randomly divided into train and test following an 80 - 20 split. This procedure was repeated independently five times, resulting in five distinct train–test partitions. Supervised approaches were fine-tuned separately on each training set, and all approaches were evaluated on the corresponding test sets. Reported results are the average performance across the five test sets.

Chunking of longer texts occurred after the split so that all chunks from the same text remain part of the same set. This prevents information leakage from the train set to the test set.

5.2 QuickUMLS

The first method tested is QuickUMLS (Soldaini and Goharian, 2016) on the task of recognition of mentions of "*Autoantibody*" and "*Disease*" entities only. One dictionary was created for each of these entity types. The dictionary for "*Disease*" was created by selecting all the UMLS terms from Semantic Type "*Disease or Syndrome*", and only keeping English strings associated with these terms. The "*Autoantibody*" dictionary was by combining the dictionary made from an intial list of HPO terms for text selection (see Subsection 3.1) and was then enriched from LOINC (Forrey et al., 1996) and manual additions. Moreover, name variations for autoantibodies were automatically added to the dictionary. QuickUMLS was used with default parameters except for the threshold, which was set to 0.9.

5.3 GLiNER

Although standard BERT-based models achieve state-of-the-art performance in NER, they rely on predefined annotation schemas and task-specific fine-tuning, in contrast to the GLiNER architecture, which supports more flexible, label-agnostic entity recognition (Zaratiana et al., 2024). GLiNER does not need a fixed list of entity types, but can work with any label passed as input at inference time. While studying a BERT model would be of interest in our use case, we decided to work with the GLiNER architecture (GLiNER 2.1 models: urchade/gliner_large-v2) on the task of NER for entity types related to autoimmunity, notably autoantibodies. Indeed, this allows for more flexibility with regards to the set of entity types to be identified, and our annotation scheme was initially not fixed. In this setup, texts longer than 384 tokens were chunked by cutting at the end of the last sentence before reaching the maximum length, and each chunk was tokenized using GLiNER's tokenizer.

GLiNER base models (small, medium, and large) were evaluated in a zero-, two-shot, and fine-tuned setting. In the GLiNER architecture, the labels are passed to the model along with the input text at inference time. The way labels are formulated may affect performances, and it may be useful to reformulate them or use synonyms to see what returns best results (e.g., "*Autoantibody target*" or "*Autoantigen*"). At this step, labels that seemed the most concise and clear were picked, but except from light changes, we did not conduct an extensive experiment to search for potential optimal alternatives. GLiNER models were evaluated on the task of NER for the following entities of interest: "*Autoantibody*", "*Autoantibody location*", "*Autoantibody target*", "*Disease*", "*Symptom or clinical sign*", on the 5 different test sets obtained from our split of the AAbAAC corpus. These entity types are the strings that were passed to GLiNER along with texts, as part of the input.

For the 2-shot setting, we passed 2 short, manually made-up examples of each entity type before the actual input text. The examples followed the format: '"sentence": "The patient was admitted after complaining of thigh pain."\n"Symptom or clinical sign": "thigh pain"'. The same 2 example sentences for each type were passed along each input text. In rare cases, the input text was long and

close to the maximal allowed chunk length, and adding the extra examples has led to truncation, which may have affected the performance slightly. For the 2-shot experiment, the GLiNER threshold parameter, which acts as a confidence filter for the model's output was set to 0.3, but for 0-shot and fine-tuned configuration it was set to 0.5.

GLiNER base models were fine-tuned using the train set from each split, and the fine-tuned models were evaluated on the associated test set. For fine-tuning, the tokenized texts of the train set are shown to the model along with their entity annotations. For evaluation, the chunked but untokenized texts of the test set are passed to the model along with the types of entities to identify.

5.4 MedGemma

Google's MedGemma model for text (google/medgemma-27b-text-it) was evaluated in 0-shot, 2-shot setting and finally fine-tuned using AAbAAC. MedGemma being a generative model, the NER task was framed as a task where the model had to output a structured JSON answer containing identified entities, as a response to a prompt containing the input text and instructions. The LLM was instructed to write a JSON string containing the text spans identified as entities. The LLM's output was then parsed to extract valid JSON segments and exclude malformed output parts. The generated answer was limited to a length of 1024 tokens. The model was not asked to return the exact character offsets delimiting the spans, as preliminary tests with this method seemed unreliable. For this experiment, the chunks of texts were given at train time, untokenized, as input to the model along with corresponding entities. For evaluation, the chunks of texts were also passed untokenized to the model along with instructions, in a prompt. The fine-tuning was conducted using rank-16 LoRA adapters, with the 4-bit quantized version of the model.

6 Results

This section presents the averaged results for the various NER methods we evaluated.

Results of the QuickUMLS experiments can be found in table Table 4. It can be observed that performances are higher for the "*Disease*" label than for the "*Autoantibody* label. This could be explained by the fact that for the recognition of "*Disease*" we relied on the richness of UMLS vocabularies: our disease dictionary is 636,989 rows, whereas our manual dictionary of "*Autoantibody*" is smaller (7,464 rows) and handcrafted, and thus probably incomplete, which would explain the lower recall for autoantibodies.

Entity type	Precision	Recall	F1 (±SD)
ALL	0.47	0.40	0.43 (±0.04)
Autoantibody	0.50	0.24	0.33 (±0.06)
Disease	0.46	0.53	0.49 (±0.06)

Table 4: Precision (P), recall (R), and F1-score (F1) of QuickUMLS for autoantibodies and diseases, using custom made dictionaries.

Table 5 reports the performance of all the different model and experimental setup we evaluated. Overall best F1-score is obtained with the fine-tuned version of MedGemma. We particularly observe that in terms of F1-score, QuickUMLS performance is not significantly lower than GLiNER and MedGemma if those are not fine-tuned. F1-score for entity type "*ALL*" is the average over 5 test sets of the micro F1-score computed by counting true positives, false positives, and false negatives of all entity types together.

Regarding the various GLiNER configurations, the fine-tuned largest version of GLiNER is overall performing slightly better than other versions. 2-shot GLiNER models perform overall worse than 0-shot, possibly indicating our entity labels were somewhat more confusing than helpful to the model without a sufficient number of examples. Top precision is achieved by GLiNER large 2-shot, whereas top recall and F1-score are achieved by GLiNER large fine-tuned (Table 5). All methods return higher precision than recall, and this seems to be especially the case in the 2-shot configurations. It should be possible to modulate this to some extent by acting on the GLiNER threshold parameter, as well as on the number and type of examples.

	Model	P	R	F1 (±SD)
	QuickUMLS	0.47	0.40	0.43 (±0.04)
0-shot	GLiNER small	0.72	0.26	0.39 (±0.05)
	GLiNER medium	0.67	0.33	0.45 (±0.04)
	GLiNER large	0.74	0.30	0.42 (±0.03)
	MedGemma	0.61	0.31	0.41 (±0.05)
2-shot	GLiNER small	0.76	0.20	0.31 (±0.05)
	GLiNER medium	0.67	0.26	0.37 (±0.04)
	GLiNER large	**0.80**	0.21	0.34 (±0.04)
	MedGemma	0.67	0.29	0.41 (±0.04)
fine-tuned	GLiNER small	0.65	0.51	0.58 (±0.05)
	GLiNER medium	0.67	0.52	0.58 (±0.03)
	GLiNER large	0.67	0.53	0.59 (±0.04)
	MedGemma	0.71	**0.62**	**0.66** (±0.03)

Table 5: Precision (P), recall (R), and F1-score (F1) of different models for all entity types together.

We also report in Table 6 per-label performances for some of the best-performing models: GLiNER large and MedGemma, both in 0-shot and fine-tuned configurations. Fine-tuning increases model F1-score for all entity types, though it sometimes slightly decreases precision: performance gains are mostly due to recall gains. Considerable performance imbalance can be observed between labels: for the fine-tuned models, the highest F1-score (0.79) is achieved by GLiNER large for the detection of "*Disease*", as per Table 6; and the lowest F1-score is achieved by MedGemma for the detection of "*Autoantibody location*" (0.33). In the case of "*Autoantibody location*", fine-tuning GLiNER large causes performance increase from an F1-score of 0.07 to 0.60; and fine-tuning MedGemma increases F1-score from 0.01 to 0.33. It is possible that some types of entities, such as "*Autoantibody location*" are not easily captured when the only information about them is their label, without examples. However, giving examples can lead to better performances, as illustrated by the performance differences between 0-shot and fine-tuned configurations.

7 Discussion

This work demonstrates the use of the AAbAAC corpus for autoimmunity information extraction through some basic NER experiments, including model fine-tuning. Indeed, models fine-tuned using the AAbAAc corpus perform better at detecting entities related to autoimmunity than the same models used out-of-the-box, or than a rule-based approach. However, so far, in the work presented here, no search for optimal fine-tuning parameters was conducted. The fine-tuned models' performances could probably be improved by performing a grid search or some other type of optimization prior to fine-tuning.

Model and Entity type	P	R	F1 (±SD)
GLiNER large 0-shot			
ALL	0.74	0.30	0.42 (±0.03)
Autoantibody	0.72	0.37	0.48 (±0.05)
Autoantibody location	0.10	0.06	0.07 (±0.05)
Autoantibody target	0.76	0.06	0.11 (±0.07)
Disease	0.82	0.48	0.60 (±0.06)
Symptom or clinical sign	0.56	0.13	0.21 (±0.06)
GLiNER large fine-tuned			
ALL	0.67	0.53	0.59 (±0.04)
Autoantibody	0.75	0.49	0.59 (±0.11)
Autoantibody location	0.47	0.88	**0.60** (±0.17)
Autoantibody target	0.43	0.28	0.34 (±0.07)
Disease	0.80	0.77	**0.79** (±0.02)
Symptom or clinical sign	0.52	0.39	**0.42** (±0.05)
MedGemma 0-shot			
ALL	0.61	0.31	0.41 (±0.05)
Autoantibody	0.65	0.45	0.53 (±0.06)
Autoantibody location	0.02	0.01	0.01 (±0.03)
Autoantibody target	0.42	0.07	0.11 (±0.02)
Disease	0.70	0.39	0.50 (±0.06)
Symptom or clinical sign	0.47	0.23	0.31 (±0.05)
MedGemma fine-tuned			
ALL	0.71	0.62	**0.66** (±0.03)
Autoantibody	0.79	0.70	**0.74** (±0.06)
Autoantibody location	0.35	0.35	0.33 (±0.18)
Autoantibody target	0.75	0.67	**0.70** (±0.04)
Disease	0.79	0.66	0.71 (±0.02)
Symptom or clinical sign	0.46	0.41	**0.42** (±0.04)

Table 6: Precision (P), recall (R), and F1-score (F1) per label for GLiNER large and MedGemma fine-tuned, both in 0-shot and fine-tuned configuration. Best F1-score per type of entity is reported in bold.

The presented AAbAAC corpus aims to be of use for complete information extraction pipelines for the study of autoimmunity. This includes both named entity recognition and extraction of relations between these entities, so the corpus also includes relation annotations. However, for now, no evaluation of the use of the corpus for information extraction was conducted. One future work may use the GLiNER2 (Zaratiana et al., 2025) architecture, which enables the joint extraction of both entities and relations simultaneously.

For GLiNER-based models, so far, only fine-tuning of the original generalist GLiNER family models was performed, however it might be of interest to attempt similar experiments with different GLiNER variants. For instance, versions dedicated to the biomedical field, such as *Ihor/gliner-biomed-large-v1.0* (Yazdani et al., 2025) are available. Attempting to fine-tune a domain-specific BERT model, such as PubMedBERT (Gu et al., 2022), for example, would also be a possible future work direction and an interesting comparison.

8 Conclusion

This work introduces AAbAAC, a manually annotated corpus of texts that was created for information extraction in the field of autoimmunity. AAbAAC includes annotations of entities and relationships that are relevant to the study of autoantibodies and autoimmune diseases. Preliminary NER experiments on this new corpus demonstrate that fine-tuning models with AAbAAC yields a performance increase over dictionary-based approaches as well as zero-shot or few-shot settings. Future work will focus on further tuning different models for this domain, for both tasks of NER and relation extraction.

9 Limitations

AAbAAC corpus is relatively small: 115 annotated texts of variable length, all of them being titles and abstracts of scientific papers published on PubMed. This is enough to deliver substantial performance gains on NER tasks for the entity types considered in this work; however, ultimately the knowledge that can be derived from this corpus is finite and may not reflect the entire existing spectrum of autoantibodies, autoimmune diseases, and the various appellations used in the literature to refer to these concepts, especially since many autoimmune diseases are rare (Miller, 2023). For some very rare, or unusually named autoantibodies not annotated in the AAbAAC corpus, detection improvements may be limited. In addition, since this corpus consists of texts drawn from the scientific literature, its impact on NER in clinical texts could be lower due to differences between written styles and possibly different naming conventions for autoantibodies.

Regarding relation annotations, some of the relation types in our scheme have almost never been instantiated in the corpus. For example, this is the case with *isCausedBy* (13 occurrences, see Table 3), a stronger relation than the more common *isAssociatedWith* (735 occurrences), or *isRefTo* (2 occurrences).

The annotation process took place using the Doccano tool, which does not allow to specify annotation attributes. Therefore, to annotate entities that are composed of several discontiguous segments of text, the *discontiguousEntity* relation is used. But this complexity is lost to an NER model that only considers entity annotations, and this may have caused some performance drop in the presented experiments.

Other models from the state of the art could have been interesting to experiment with as well, in particular BERT-type models fine-tuned for the biomedical literature, or non-proprietary LLM. Their inclusion could open perspectives to extend our study to consider the computational cost of various approaches and attempt to specialize or distill efficient but costly models.

Acknowledgments

The authors acknowledge the Filières de Santé Maladies Rares BRAIN-TEAM and Filnemus for funding. The authors thank B. Belloir and S. Bernichtein from BRAIN-TEAM and R. Soussi from Filnemus for their help. The authors highlight the contributions of the BRAIN-TEAM, Filnemus, FAI²R, Fimarad, Fimatho, FILFOIE, MHEMO, and DéfiScience networks. The authors also thank C. Lucano and C. Fabrizzi (ORPHANET) for their help and our research teams, HeKA and Pathophysiological basis of skeletal dysplasia, for their constant support.

References

Haseeb Ahsan. 2023. Origins and history of autoimmunity—A brief review. *Rheumatology & Autoimmunity*, 3(1):9–14.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Thuy Trang Dinh, Trang Phuong Vo-Chanh, Chau Nguyen, Viet Quoc Huynh, Nam Vo, and Hoang Duc Nguyen. 2022. Extract antibody and antigen names from biomedical literature. *BMC Bioinformatics*, 23:524.

A. W. Forrey, C. J. McDonald, G. DeMoor, S. M. Huff, D. Leavelle, D. Leland, T. Fiers, L. Charles, B. Griffin, F. Stalling, A. Tullis, K. Hutchins, and J. Baenziger. 1996. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, 42(1):81–90.

Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M. Bagley, Eduard Bakštein, James P. Balhoff, Gareth Baynam, Susan M. Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F. Bodenstein, Pablo Botas, Kaan Boztug, and 157 others. 2024. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Scott M. Hayter and Matthew C. Cook. 2012. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmunity Reviews*, 11(10):754–765.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *Open Journal of Modern Linguistics*, 9(3):206–214.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Frederick W. Miller. 2023. The increasing prevalence of autoimmunity and autoimmune diseases: An urgent call to action for improved understanding, diagnosis, treatment and prevention. *Current opinion in immunology*, 80(102266).

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with UMLS concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808.

Adam Remaki, Jacques Ung, Pierre Pages, Perceval Wajsburt, Elise Liu, Guillaume Faure, Thomas Petit-Jean, Xavier Tannier, and Christel Gérardin. 2025. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study. *JMIR Medical Informatics*, 13:e68704.

Eric Sayers. 2018. E-utilities quick start. In *Entrez® Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

Yinon Shapira, Nancy Agmon-Levin, and Yehuda Shoenfeld. 2010. Defining and analyzing geoepidemiology and human autoimmunity. *Journal of Autoimmunity*, 34(3):J168–J177.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Sandeep Subramanian and Madhavi K. Ganapathiraju. 2017. Antibody exchange: Information extraction of biological antibody donation and a web-portal to find donors and seekers. *Data*, 2(4):38.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in python. *Journal of Open Source Software*, 6(62):2989.

Randi Vita, Nina Blazeska, Daniel Marrama, IEDB Curation Team Members, Sebastian Duesing, Jason Bennett, Jason Greenbaum, Marcus De Almeida Mendes, Jarjapu Mahita, Daniel K. Wheeler, Jason R. Cantrell, James A. Overton, Darren A. Natale, Alessandro Sette, and Bjoern Peters. 2025. The immune epitope database (IEDB): 2024 update. *Nucleic Acids Research*, 53:D436–D443.

Amy Y. Wang, Jeremiah H. Sable, and Kent A. Spackman. 2002. The SNOMED clinical terms development process: refinement and analysis of content. *Proceedings of the AMIA Symposium*, pages 845–849.

Dan Wang, Liuhui Yang, Ping Zhang, Joshua LaBaer, Henning Hermjakob, Dong Li, and Xiaobo Yu. 2017. AAgAtlas 1.0: a human autoantigen database. *Nucleic Acids Research*, 45(D1):D769–D776.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: Schema-driven multi-task learning for structured information extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 130–140. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

What Do Biomedical NER and Entity Linking Benchmarks Measure? A Corpus-Centric Diagnostic Framework

Robert Leaman* **Rezarta Islamaj*** **Zhiyong Lu**
National Library of Medicine, Bethesda, MD
Robert.Leaman@nih.gov Rezarta.Islamaj@nih.gov Zhiyong.Lu@nih.gov

Abstract

Biomedical named entity recognition (NER) and entity linking (EL) strongly depend on annotated corpora, but the utility of these resources for benchmarking is often assumed rather than characterized. We present a corpus-centric framework for diagnosing benchmark-relevant properties directly from corpus annotations, concept links, train-test splits, document metadata, and terminology mappings. The framework organizes standardized statistics into five families: (1) scale, density and label distribution, (2) lexical and conceptual structure, (3) train-test overlap, (4) metadata composition, and (5) terminology coverage where applicable. Applying the framework to nine corpora spanning diseases, chemicals, and cell types, we find that corpus properties can differ substantially, even when they address the same apparent task. We find differences in the evaluation signal they provide, the generalization demands they impose, the degree of train–test reuse they permit, and the regions of biomedical literature and concept space they represent. These differences suggest that commonly reported corpus statistics can be insufficient to characterize what biomedical NER and EL benchmarks evaluate. We argue that corpus-centric diagnostics provide a practical framework for analyzing corpora beyond surface descriptors such as corpus size and entity type, for identifying potential transfer risks, and for interpreting the scope of benchmarking conclusions. We release the framework as open-source code[1] with an interactive dashboard to support reproducing our analyses and characterizing additional corpora.

1 Introduction

Extracting structured information from biomedical literature requires systems to identify and link entities (e.g., genes, diseases, chemicals) to standardized identifiers. These grounding tasks—named entity recognition (NER) and entity linking (EL)—remain critical in the large language models (LLM) era to ensure outputs are auditable, comparable, and reusable.

Manually annotated corpora serve both as training data and as evaluation benchmarks (Kim et al., 2003; Collier et al., 2004; Morgan et al., 2008; Lu et al., 2011; Wei et al., 2013; Doğan et al., 2014; Krallinger et al., 2015; Li et al., 2016; Islamaj et al., 2021b,a; Bada et al., 2012; Herrero-Zazo et al., 2013; Wei et al., 2016; Miranda-Escalada et al., 2023). When used as benchmarks, they function as measurement instruments: the key question is not only whether annotations are correct, but what capabilities the benchmark tests and whether conclusions transfer to intended use cases.

This distinction matters because benchmark utility is task- and domain-dependent. A corpus can be carefully annotated yet too narrow, homogeneous, or leaky across splits to support informative evaluation. Biomedical natural language processing (NLP) values rare, specialized, and emerging concepts, making evaluation sensitive to which entities, subdomains, time periods, and document types are represented. Without characterizing the *corpus domain* and the target *application domain*, it is difficult to distinguish generalization issues from corpus-specific artifacts, leakage, or domain mismatch. Benchmark corpora have primarily been compared by size, entity type, or reported system performance. These descriptors do not reflect overlap risks, lexical difficulty, domain bias, or concept coverage.

To address this gap, we introduce a corpus-centric framework that computes standardized statistics directly from annotations, concept links, corpus splits, metadata, and terminologies (Figure 1). These statistics provide a multidimensional analysis for PubMed- and PMC-based NER and EL corpora: density indicates how much evaluation signal is available; lexical and conceptual variation

* These authors contributed equally
[1] https://github.com/NLM-DIR/CorpusBenchmarking

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 801–811
July 3-4, 2026 ©2026 Association for Computational Linguistics

indicate what kinds of generalization are required; overlap reveals leakage risk; metadata composition characterizes the literature represented; and terminology coverage indicates which parts of the concept space are represented.

Applied to nine corpora spanning diseases, chemicals, and cell types, the framework shows that resources similar by task label or size can differ substantially across these signals. Our contributions are: (1) a corpus-as-instrument framing for biomedical NER and EL benchmarks; (2) a practical framework of corpus-centric diagnostics; (3) an analysis showing how structural differences inform evaluation sensitivity, leakage risk, coverage, and transferability; and (4) open-source code and an interactive dashboard for reproducing our results and for analyzing new corpora.

2 Related Work

Biomedical entity annotation corpora have proliferated over the years, yet they are routinely treated as benchmarks without any systematic analysis of what they actually measure. Early efforts, such as GENIA (Kim et al., 2003) established large-scale manual mention annotation with fine-grained semantic categories, later simplified for shared tasks such as JNLPBA (Collier et al., 2004). Subsequent corpora introduced entity normalization, typically targeting a single entity type: diseases (NCBI Disease), chemicals (CHEMDNER, BC5CDR), genomic variants (tmVar) and genes (BioCreative challenges) (Doğan et al., 2014; Krallinger et al., 2015; Li et al., 2016; Wei et al., 2013; Morgan et al., 2008). Resources such as CRAFT (Bada et al., 2012) broadened the scope to multi-entity, full-text annotation. More recent corpora—including NLM-Chem, BioRED, and CellLink–have further expanded document and entity coverage, and incorporate richer annotation structures such as relations (Islamaj et al., 2021b, 2024; Rotenberg et al., 2026). Despite being widely used together, these corpora vary considerably in document type, annotation density, normalization support, temporal range, and domain focus—differences that are rarely examined in terms of what each benchmark actually evaluates.

NLP research has shown that dataset properties can distort benchmark interpretation. Work on saturation motivates aggregated benchmarks such as GLUE and SuperGLUE (Wang et al., 2018, 2019); work on artifacts, leakage, and memorization shows that apparent improvements can reflect shortcuts or overlap rather than the intended capability (Gururangan et al., 2018; Liang et al., 2023; Tutubalina et al., 2020); and HELM emphasizes multi-metric evaluation (Liang et al., 2023). Biomedical suites such as BLUE, BLURB, and BigBIO standardize cross-task evaluation (Peng et al., 2019; Gu et al., 2021; Fries et al., 2022), but generally treat corpora as fixed inputs rather than explaining how corpus properties shape validity or transferability.

Annotation quality measures, particularly inter-annotator agreement (IAA), assess consistency but not benchmark scope. Prior work distinguishes agreement across span boundaries, labels, and concept links (Artstein and Poesio, 2008), recommends F1 for span-based tasks lacking a well-defined negative class (Hripcsak and Rothschild, 2005), and notes that disagreement may reflect ambiguity, error, or guideline limitations (Aroyo and Welty, 2015; Uma et al., 2021). High agreement is necessary but not sufficient: simplifying annotation can increase agreement while removing realistic ambiguity (Hovy and Lavid, 2010). Corpus papers often report counts and distributions, but these statistics are rarely organized around evaluation claims. Our framework connects such descriptions to overlap, memorization, domain shift, annotation scope, and terminology coverage.

3 Methods

3.1 Framework Overview and Representation

Our framework characterizes NER and EL corpora through four stages: conversion to a shared representation, filtering, metric computation, and visualization. Corpora are standardized into documents containing text, optional metadata, and annotations (spans, surface forms, labels, and linked concept identifiers).

This design supports both NER-only and NER+EL PubMed- or PMC-based corpora. NER-only corpora are evaluated on text-, span-, and mention-level statistics, while EL-supported datasets additionally yield concept-level diagnostics. Metrics are computed over configurable corpus bundles, comparison suites, entity scopes, and train/dev/test partitions to enable interpretable comparisons across heterogeneous datasets.

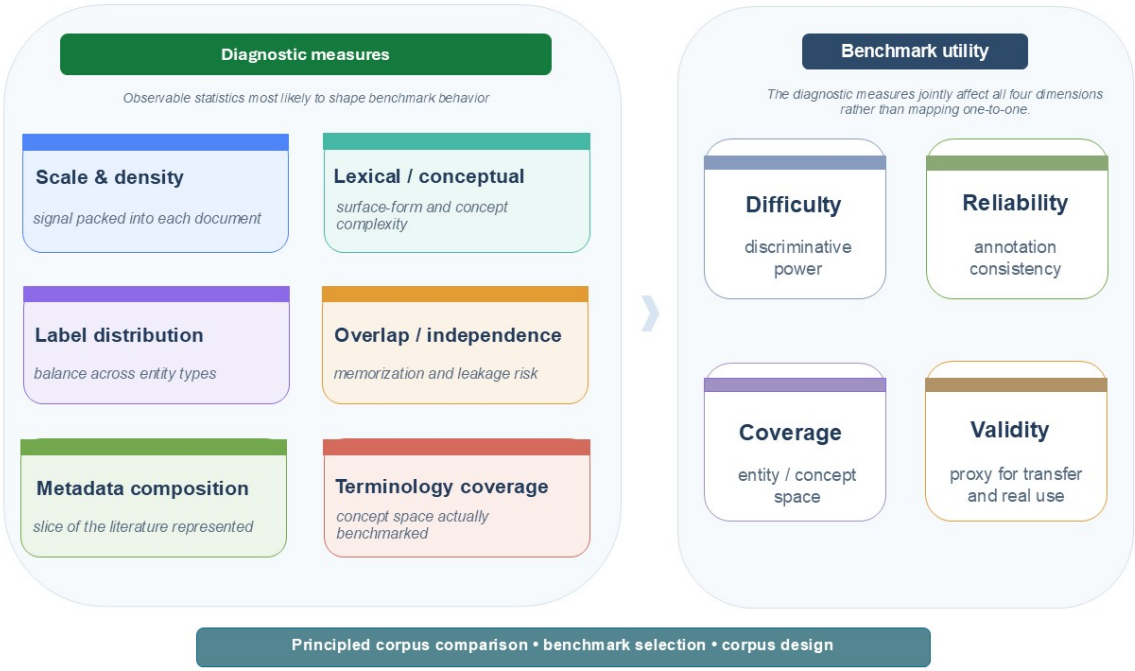


Figure 1: Corpus diagnostic framework. Entity-annotated corpora are converted into a common representation, enabling computation of statistics over annotations, identifiers, and metadata. These statistics characterize scale and density, lexical and conceptual variation, train-test overlap, metadata composition, and terminology coverage, enabling principled comparison of biomedical NER and EL benchmarks.

3.2 Corpora Analyzed

We apply the framework to nine corpora spanning diverse entity types (e.g., diseases, chemicals, cell types) and document scopes (abstracts, captions, full-text): AnatEM (Pyysalo and Ananiadou, 2014), BC5CDR (Li et al., 2016), BioID (Arighi et al., 2017), CHEMDNER (Krallinger et al., 2015), CRAFT (Bada et al., 2012), CellLink (Rotenberg et al., 2026), JNLPBA (Collier et al., 2004), NCBI-Disease (Doğan et al., 2014), and NLM-Chem (Islamaj et al., 2021b). Where public test data were unavailable or original annotation layers were altered, we utilized the closest documented subset and note these limitations alongside the relevant results.

3.3 Diagnostic Statistics

The framework computes corpus statistics across five families to diagnose benchmark properties prior to system evaluation:

- **Scale, Density, and Label Distribution**: We compute total documents, tokens, annotations, and unique mentions/identifiers per document.
- **Lexical and Conceptual Structure**: For normalized corpora, we measure mention ambiguity (the number of distinct label/link pairs mapped to a single surface form) and identifier variation (the number of distinct surface forms mapped to a label/link pair). These distinguish a benchmark's demand for contextual disambiguation versus its demand for recognizing diverse lexical realizations.
- **Train-Test Overlap**: To assess leakage and memorization risk, we compute Jaccard overlap between train and test splits at four abstraction levels: general token vocabulary, tokens inside entity mentions, exact mention strings, and concept identifiers.
- **Metadata Composition**: We profile the represented literature slice via temporal statistics (publication year ranges and distributions) and journal diversity (unique journals and top-journal concentration). We derive broad topic profiles from article Medical Subject Headings (MeSH) (Lipscomb, 2000) topics where available, falling back to NLM Catalog MeSH journal topics if necessary.
- **Terminology-Aware Coverage**: For diseases and chemicals, we link identifiers to their respective concepts in MeSH, MONDO (Vasilevsky et al., 2026), or

ChEBI (Malik et al., 2026); for cell types, we support Cell Ontology (CL) (Tan et al., 2026). We quantify vocabulary coverage by analyzing the distribution of annotations across high-level branches and compute hierarchy depth as a proxy for concept specificity.

3.4 Implementation

The framework is implemented as an open-source, YAML-configurable Python pipeline that outputs structured JSON statistics. It includes acquisition specifications for downloading, extracting, converting, caching, and validating expected corpus files. Input support includes BioC XML (Comeau et al., 2013, 2019), PubTator (Wei et al., 2024), BRAT/standoff (Stenetorp et al., 2012), and Knowtator (Ogren, 2006) annotations, with registry-based extension points for additional loaders and metrics. Ontologies in OBO format are directly supported as terminology sources. The accompanying self-contained HTML/JavaScript dashboard[2] combines scale, overlap, metadata, terminology, and entity-scope views to reproduce our analyses and evaluate new corpora.

4 Results

We use the nine corpora to illustrate how corpus-centric diagnostics clarify the evaluation role of a benchmark. The goal is not to rank corpora, but to demonstrate that datasets with similar task labels often function as different measurement instruments: they expose systems to different volumes of evaluation signal, different forms of lexical and conceptual generalization, different leakage risks, and different regions of the biomedical literature and concept space.

4.1 Scale, annotation density, lexical and conceptual variation

Table 1 reports statistics for nine heterogeneous biomedical corpora. These measurements show that corpora differ fundamentally in the structural nature of the evaluation signal they provide. Annotation density varies widely across corpora, reflecting differences in document scope, annotation unit, and entity scope. Dense full-text corpora such as NLM-Chem and CRAFT provide many labeled decisions per article, whereas passage-, abstract-, and caption-based corpora distribute fewer decisions across more sampled text units. This raw density is useful because it affects the number of system decisions contributing to an evaluation estimate. However, annotations from the same full-text article are not necessarily independent: repeated mentions, recurring identifiers, and section-specific language can increase decision volume without proportionally increasing lexical, conceptual, or contextual diversity. Density should therefore be interpreted as signal concentration, not as a direct measure of task difficulty or benchmark quality. In this sense, full-text corpora evaluate behavior over long-document contexts and repeated real-world usage, while shorter-unit corpora may provide broader sampling of independent contexts per annotation. Concept-level diagnostics further define what the instrument is calibrated to measure. Variation ranges from 1.48 surface forms per label/link pair in BioID to 3.74 in CellLink, indicating differing demands on lexical generalization.

4.2 Train-Test Overlap

Figure 2 reports train-test Jaccard overlap at four levels: token vocabulary, mention tokens, exact mention strings, and concept identifiers. These levels distinguish increasingly task-specific forms of reuse. Across all corpora, token vocabulary overlap is higher than mention-token overlap, which is, in turn, higher than exact mention string overlap.

The slope of this drop varies and is itself informative. JNLPBA falls from 35.9% token overlap to 6.6% mention-string overlap, indicating that its test split contains relatively little exact entity-form reuse. CellLink presents a different structural pattern: its mention-token overlap (27.9%) is much higher than its exact mention-string overlap (9.1%), while identifier overlap is higher still (40.6%). This profile suggests compositional novelty, where new cell-population names are constructed from familiar component tokens and mapped to familiar concepts. Examples include "CD4+ T cells" (CL:0000624), "resting CD4 memory T cells" (CL:0000897), and "resting NK cells" (CL:0000623), which combine recognizable modifiers with cell-type heads.

Identifier overlap adds a concept-level view unavailable from strings alone. Among normalized corpora, identifier overlap ranges from 14.2% in BioID to 40.6% in CellLink—a range wider than the corresponding mention string overlaps, and with a different ordering. CellLink's high identifier overlap paired with low string overlap means it tests the ability to map novel text to relatively fa-

[2] https://nlm-dir.github.io/CorpusBenchmarking/dashboard.html

Corpus	Doc. type	Docs	Tokens	E	Total Ann.	Ann/ doc	Men./ doc	IDs/ doc	ID vocab	Ambi-guity[a]	Vari-ation[b]
AnatEM	abstracts	1,212	259,510	12	13,701	11.3	7.2	—	—	—	—
BC5CDR	abstracts	1,500	297,019	2	29,271	19.5	9.4	6.9	MeSH	1.02	2.47
BioID	captions	13,697	771,248	8	102,742	7.5	4.9	5.2	mixed[c]	1.35	1.48
CHEMDNER	abstracts	10,000	2,092,491	1	84,331	8.4	4.6	—	—	—	—
CRAFT	full text	97	652,168	11	99,623	1027.0	241.9	149.4	mixed[c]	1.02	2.33
CellLink	passages	2,003	227,490	3	14,731	7.3	6.0	4.1	CL	1.07	3.74
JNLPBA	abstracts	2,404	564,660	5	59,963	24.9	16.5	—	—	—	—
NCBI-Disease	abstracts	793	169,561	1	6,892	8.7	5.1	3.2	MeSH/OMIM	1.01	2.75
NLM-Chem	full text	150	789,532	1	38,339	255.6	54.3	34.1	MeSH	1.02	2.42

[a] **Ambiguity**: mean label/link pairs per unique mention string. Values near 1.0 indicate most mentions map to a single label/link pair. [b] **Variation**: mean surface forms per label/link pair. Reported only for corpora with concept-level identifiers; AnatEM, CHEMDNER, and JNLPBA are excluded (n/a). [c] BioID and CRAFT link entity types to many ontology identifier resources.

Table 1: Basic statistics for each corpus. E = number of annotated entity types; Men./doc = unique mention strings per document; IDs/doc = unique concept identifiers per document. CellLink values reflect the currently released annotated train and development partitions.

miliar concepts. Conversely, NCBI-Disease's low identifier overlap requires systems to generalize to genuinely novel conceptual territory.

4.3 Metadata Composition

Temporal coverage differs sharply: CellLink is most recent, spanning 2019-2025; CHEMDNER is nearly a single-year snapshot (97.7% from 2013); CRAFT spans 2001-2007; NCBI-Disease is predominantly pre-2000 (90.8%); and BC5CDR is the broadest, spanning 1968-2016.

The journal distribution shows similar variation in domain. BC5CDR is by far the most venue-diverse: 703 journals, with the top five journals accounting for only 5.0% of the corpus. BioID represents the opposite extreme, with 18 journals total and 83.2% of corpus documents from the top five journals.

The article-topic mappings in Table 2 demonstrate substantial differences in the slice of biomedical literature these corpora represent. NCBI-Disease is strongest in genetics/genomics, CHEMDNER and NLM-Chem emphasize chemistry/materials, BioID is dominated by molecular biology/biochemistry, and CellLink is enriched for general biology and cell/developmental biology.

4.4 Terminology-based coverage

Figure 3 demonstrates that corpora sharing an entity label may still exercise different concept regions of the concept space. The left panels normalize branch counts within each corpus, showing what share of the corpus's own annotations falls into each high-level terminology branch. The right panels normalize the same branch counts by the size of the corresponding terminology branch, highlighting how the corpus represents different areas of the terminology. Figure 3 focuses on MeSH diseases and chemicals; analogous CL-based statistics for cell-type corpora are available in the dashboard.

For diseases, NCBI-Disease is dominated by Congenital and Hereditary Diseases (C16, 23%) and Neoplasms (C04, 14%), consistent with its genetics origin. BC5CDR instead peaks at Pathological Conditions (C23, 20%) and Cardiovascular Diseases (C14, 17%), consistent with pharmacovigilance as a clinically cross-cutting focus rather than a narrow specialty focus. Chemically-Induced Disorders account for 5.8% of BC5CDR disease annotations but only 0.05% of NCBI-Disease. Thus, while both corpora evaluate "disease" recognition, they measure fundamentally distinct capabilities.

For chemicals, BC5CDR is concentrated in Organic Chemicals (D02, 37%) and Heterocyclic Compounds (D03, 28%), the categories containing most small-molecule drugs. NLM-Chem is more broadly distributed, with higher representation of Inorganic Chemicals (D01, 18%) and Amino Acids/Proteins (D12, 12%).

5 Discussion

The central implication of this study is that benchmark results for biomedical NER and EL cannot be interpreted independently of the corpora that produced them. A benchmark corpus is not simply a labeled sample of text: it is a measurement instrument whose structure determines which system capabilities are exercised and limits how far evaluation conclusions can reasonably transfer. Because benchmark utility is task- and domain-dependent,

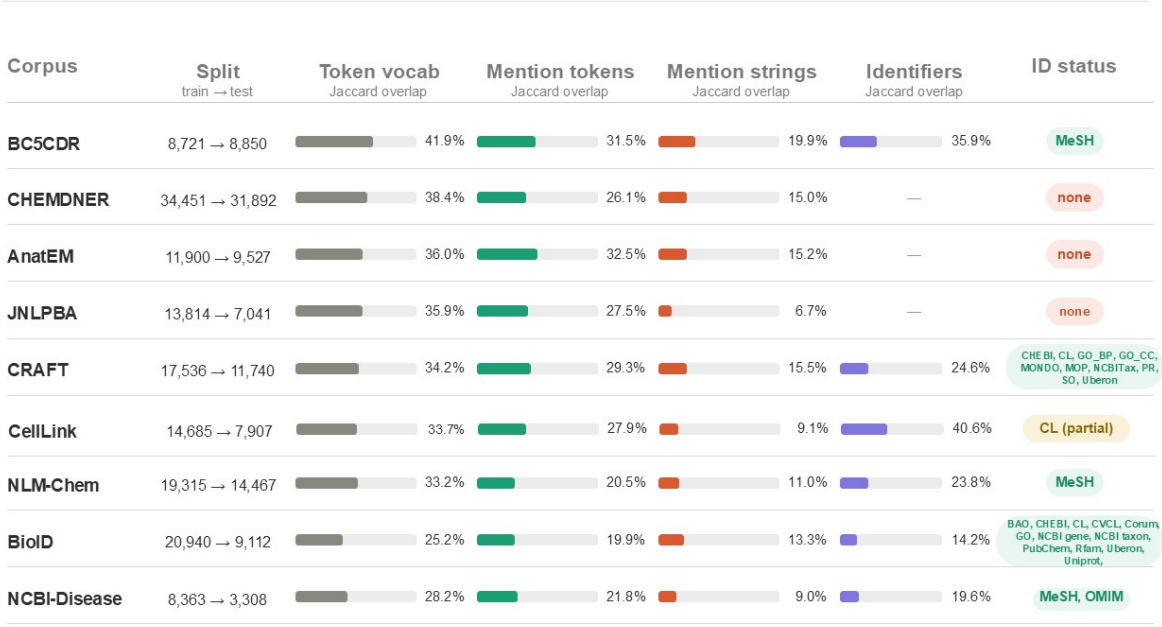


Figure 2: Train-test overlap across nine biomedical corpora. All values are Jaccard similarity (%) between training and test splits. Token vocab: unique vocabulary tokens shared between splits. Mention tokens: unique tokens within entity mention spans. Mention strings: exact entity surface forms shared between splits. Identifiers: concept identifiers shared between splits.

resources that share an entity type label can still function as fundamentally different instruments, evaluating different capabilities, imposing different generalization demands, and representing different regions of biomedical literature and concept space. No single statistic captures this multidimensionality. Instead, benchmark interpretation depends on the interaction of several diagnostic signals, and the nine corpora examined here illustrate how these signals combine in practice.

Annotation density and label distribution. Annotation density determines how concentrated evaluation signal is within each sampled unit. Dense full-text corpora such as CRAFT and NLM-Chem yield many linked decisions from relatively few articles, which can improve evaluation sensitivity and expose systems to full-text sectional variation. At the same time, raw density is partly confounded with document length and annotation scope. A large number of annotations from one article may include repeated mentions of the same entities, and therefore may not provide the same independent evidence as the same number of annotations sampled from many distinct passages or documents. Sparse corpora based on abstracts, passages, or figure captions require broader sampling to achieve stable estimates, and they provide more independent contexts per annotation. Each design serves different evaluation purposes: full-text corpora emphasize document-level realism and repeated-use behavior, whereas passage- or abstract-based corpora can emphasize breadth of contexts under fixed annotation budgets.

Lexical and conceptual variation. Variation determines whether a corpus tests recognition of repeated forms, alternative expressions for known concepts, or generalization to genuinely new ones. CellLink's variation of 3.74 surface forms per label-link pair reflects the compositional naming conventions of single-cell biology: novel cell-population names are assembled from familiar modifiers and cell-type heads, so systems must generalize across surface forms while mapping to relatively familiar concepts. BC5CDR's lower variation is consistent with the pharmacovigilance literature, where drug and disease names tend to be standardized.

Train-test overlap. Overlap analysis offers a particularly direct diagnostic for interpreting unexpectedly strong or weak results, consistent with prior work on leakage and memorization (Tutubalina et al., 2020; Liang et al., 2023). The steepness of the token-level overlap drop varies and is in itself informative. CellLink illustrates one diagnostic profile: high identifier overlap (40.6%)

Topic	AnatEM	BC5CDR	BioID	CHEM	CRAFT	CellLink	JNLPBA	NCBI	NLM
Molecular bio. / biochemistry	**16%**	6%	**62%**	15%	21%	12%	**36%**	17%	18%
Chemistry / materials science	8%	15%	—	**29%**	4%	4%	11%	5%	**29%**
Genetics/genomics	5%	—	8%	3%	18%	8%	10%	**28%**	4%
General bio. / anatomy / physiology	14%	14%	6%	18%	**23%**	**22%**	11%	11%	14%
Demographic characteristics	7%	**18%**	—	4%	4%	5%	3%	16%	6%
Cell & developmental biology	7%	1%	10%	5%	9%	17%	12%	4%	6%
Clinical specialties by organ system	8%	11%	—	3%	7%	9%	2%	5%	4%
General / internal medicine	6%	6%	7%	2%	1%	3%	0%	1%	2%
Pharmacology	2%	6%	—	6%	—	—	1%	—	2%
Other	27%	23%	7%	15%	13%	20%	14%	13%	15%

Topics are assigned from article MeSH terms where available; unresolved article terms fall back to NLM Catalog MeSH journal topics and configured journal-name anchors. "Other" aggregates all categories not shown. The largest displayed non-Other entry in each column is shown in bold.

Table 2: Broad research topic composition of each corpus. Values represent approximate percentage of corpus articles from each topic area. CHEM = CHEMDNER; NCBI = NCBI-Disease; NLM = NLM-Chem.

combined with much lower mention-string overlap (9.1%) means the benchmark emphasizes compositional surface-form generalization rather than novel concept recognition. NCBI-Disease presents the opposite profile, with low identifier overlap (19.6%) that places genuine demands on concept-level generalization. JNLPBA's sharp drop from 35.9% token overlap to 6.7% mention-string overlap, combined with the absence of normalization, may partially recontextualize its relatively low performance ceiling [e.g., (Huang et al., 2020)]: systems cannot rely heavily on exact entity-form reuse.

Metadata composition. Temporal distributions, journal diversity, and topic profiles describe the slice of literature represented by the corpus. These properties define the domain over which benchmark claims can reasonably apply and determine whether conclusions transfer to corpora drawn from different time periods, venues, or subfields. CHEMDNER's near-total concentration in 2013 chemistry literature and BioID's extreme journal concentration (83.2% from five journals) are structural features that may limit the generalizability of results. BC5CDR's breadth across 703 journals and five decades makes performance claims more domain-general. These metadata properties are rarely reported yet have a bearing on whether performance on a benchmark supports claims about a target deployment setting.

Terminology coverage. The contrast between NCBI-Disease and BC5CDR illustrates why terminology-aware analysis matters even when two corpora nominally share a task label. Both annotate disease entities, yet NCBI-Disease is dominated by Congenital and Hereditary Diseases (C16) and Neoplasms (C04), reflecting its genetics origin, while BC5CDR peaks at Pathological Conditions (C23) and Cardiovascular Diseases (C14), reflecting its pharmacovigilance focus. Chemically-Induced Disorders (C25) account for 5.8% of BC5CDR disease annotations but only 0.05% of NCBI-Disease. High performance on one therefore supports only limited claims about transfer to the other, despite the shared entity label. The same logic applies to chemical corpora: BC5CDR's concentration in Organic Chemicals and Heterocyclic Compounds reflects a small-molecule drug emphasis, while NLM-Chem's broader distribution across Inorganic Chemicals, Amino Acids, and Lipids reflects the wider biochemical scope of full-text molecular biology literature. Differences in terminology coverage may reflect different practical capabilities even when their reported task names are identical.

From diagnostics to corpus decisions. Corpus diagnostics are informative not only retrospectively but before a benchmark is finalized. Before releasing a split, developers can compute mention-string and identifier overlap to detect leakage. While extending a corpus, they can sample documents to fill temporal, journal, topic, or terminology-branch gaps. When selecting a benchmark, researchers can choose corpora whose density, overlap profile, and terminology coverage match the intended deployment claim. When combining corpora, they can verify whether a new source adds concept regions not already represented or mainly duplicates familiar strings and identifiers. The goal, then, is to tie benchmark interpretation to corpus characteristics and intended use, rather than assign a single quality score.

Limitations. The framework characterizes corpus structure but does not directly predict down-

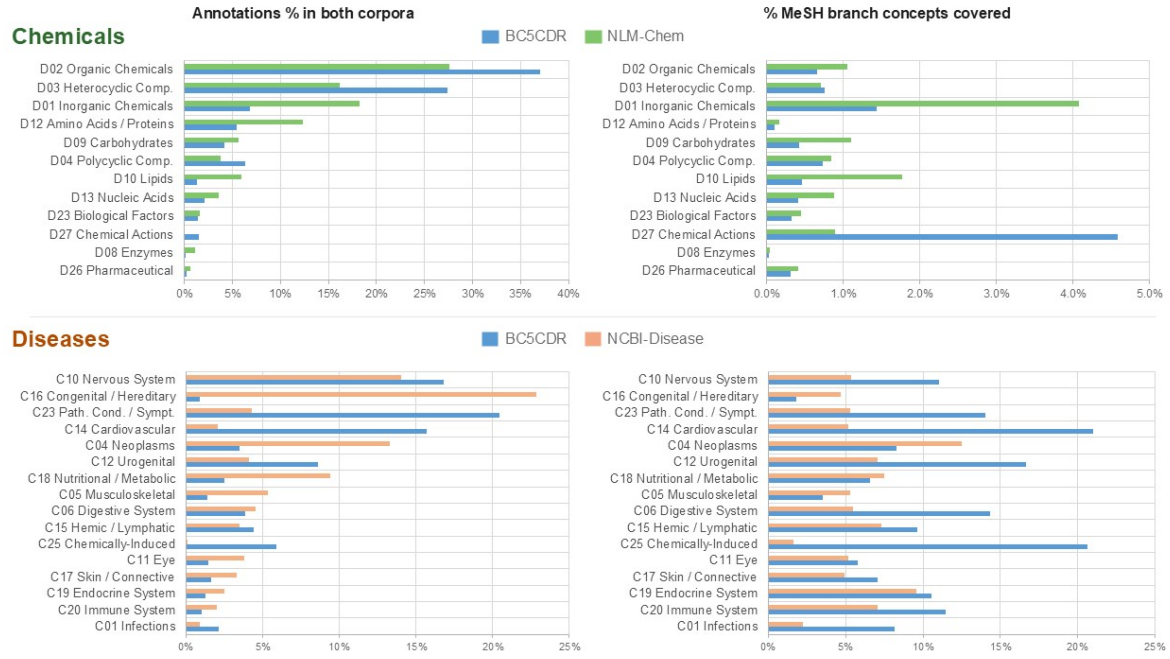


Figure 3: Terminology-based coverage analysis. Left panels: distribution normalized within each corpus's own annotations. Right panels: distribution normalized within the MeSH tree vocabulary. Top panels: distribution across MeSH chemical branches (D-branches) for BC5CDR and NLM-Chem. Bottom panels: distribution across MeSH disease branches (C-branches) for BC5CDR and NCBI-Disease.

stream system rankings or benchmark saturation. Measured ambiguity conflates true polysemy, underspecified guidelines, and annotation errors; the present framework does not distinguish among these sources. Surface-form variation measures observed diversity rather than the full range of synonymy present in the literature. Terminology analyses depend on the version of the reference ontology used. Finally, topic mappings provide practical, lightweight domain diagnostics rather than definitive subject classifications.

6 Conclusion

We presented a corpus-centric framework for diagnosing the benchmark utility of biomedical corpora with entity annotations. The framework organizes standardized statistics over annotations, linked identifiers, corpus splits, document metadata, and terminology mappings into five diagnostic families: (1) scale, density and label distribution, (2) lexical and conceptual structure, (3) train-test overlap, (4) metadata composition, and (5) terminology coverage. Applied to nine biomedical NER and EL corpora, these diagnostics show that corpora for the same apparent task can measure substantially different capabilities.

The main implication is that these statistics provide a practical basis for more complete reporting of biomedical NER and EL benchmarks. Corpus reports should describe annotation density, normalization support, lexical and identifier variation, train-test overlap at multiple abstraction levels, temporal and journal composition, and terminology coverage where applicable. Without these measurements, benchmark results are difficult to interpret. Standardized corpus diagnostics would make evaluation claims more interpretable, clarify what a benchmark measures, and support transparent, reproducible, transfer-aware evaluation.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The NIH author contributions are considered Works of the United States Government. The findings and conclusions presented are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, Samuel Bayer, Robin Liechti, Donald Comeau, and

Cathy Wu. 2017. Bio-ID track overview. In *BioCreative VI Challenge Evaluation Workshop*, pages 14–19, Bethesda, MD.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner Jr, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13:161.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, 2013:bat064.

Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. 2019. PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, 35(18):3533–3535.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, and 1 others. 2022. BigBio: A framework for data-centric biomedical natural language processing. In *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1–23.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data, New Orleans, Louisiana, Association for Computational Linguistics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Ming-Siang Huang, Po-Ting Lai, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2020. Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, 21(6):2219–2238.

Islamaj, R., C. H. Wei, D. Cissel, N. Miliaras, O. Printseva, O. Rodionov, K. Sekiya, J. Ward, and Z. Lu. 2021a. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of Biomedical Informatics*, 118:103779.

Islamaj, R., C. H. Wei, P. T. Lai, L. Luo, C. Coss, P. Gokal Kochar, N. Miliaras, O. Rodionov, K. Sekiya, D. Trinh, D. Whitman, and Z. Lu. 2024. The biomedical relationship corpus of the BioRED track at the BioCreative VIII challenge and workshop. *Database (Oxford)*, 2024:baae099.

Rezarta Islamaj, Robert Leaman, Sun Kim, Dongseop Kwon, Chih-Hsuan Wei, Donald C. Comeau, Yifan Peng, David Cissel, Cathleen Coss, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Stella Koppel, Dorothy Trinh, Keiko Sekiya, Janice Ward, Deborah Whitman, Susan Schmidt, and Zhiyong Lu. 2021b. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(1):91.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–i182.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, Roger A Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos,

David Campos, Buzhou Tang, Hua Xu, and 34 others. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266.

Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, Hye-Cheol Cho, Martin Gerner, Illes Solt, Shashank Agarwal, Feifan Liu, Dina Vishnyakova, Patrick Ruch, Martin Romacker, Fabio Rinaldi, and 9 others. 2011. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12:S2.

Adnan Malik, Muhammad Arsalan, Carlos Moreno, Juan Mosquera, Eloy Félix, Tevfik Kizilören, Venkatesh Muthukrishnan, Barbara Zdrazil, Andrew R Leach, and Noel M O'Boyle. 2026. ChEBI: re-engineered for a sustainable future. *Nucleic Acids Research*, 54:D1768–D1778.

Miranda-Escalada, A., F. Mehryary, J. Luoma, D. Estrada-Zavala, L. Gasco, S. Pyysalo, A. Valencia, and M. Krallinger. 2023. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical-protein relations. *Database (Oxford)*, 2023:baad080.

Alexander A. Morgan, Zhiyong Lu, Xinglong Wang, Aaron M. Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, Chenjie Sun, Heng-hui Liu, Rafael Torres, Michael Krauthammer, William W. Lau, Hongfang Liu, Chun-Nan Hsu, Martijn Schuemie, K. Bretonnel Cohen, and Lynette Hirschman. 2008. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3.

Philip V. Ogren. 2006. Knowtator: A protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 273–275, New York City, USA. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Rotenberg, N. H., R. Leaman, R. Islamaj, H. Kuivaniemi, G. Tromp, B. Fluharty, S. Richardson, C. Eastwood, M. Diller, B. Xu, A. V. Pankajam, D. Osumi-Sutherland, Z. Lu, and R. H. Scheuermann. 2026. Cell phenotypes in the biomedical literature: a systematic analysis and text mining corpus. *bioRxiv*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Shawn Zheng Kai Tan, Aleix Puig-Barbe, Damien Goutte-Gattat, Caroline Eastwood, Brian Aevermann, Alida Avola, James P. Balhoff, Ismail Ugur Bayindir, Jasmine Belfiore, Anita Reane Caron, David S. Fischer, Nancy George, Benjamin M. Gyori, Melissa A. Haendel, Charles Tapley Hoyt, Huseyin Kir, Tiago Lubiana, Nicolas Matentzoglu, James A. Overton, and 16 others. 2026. The cell ontology in the age of single-cell omics. *Scientific Data*.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for BERT-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Nicole A Vasilevsky, Sabrina Toro, Nicolas Matentzoglu, Joseph E Flack, Kathleen R Mullen, Harshad Hegde, Sarah Gehrke, Patricia L Whetzel, Yousif Shwetar, Nomi L Harris, Mee S Ngu, Gioconda L Alyea, Megan S Kane, Paola Roncaglia, Eric Sid, Courtney L Thaxton, Valerie Wood, Roshini S Abraham, Maria Isabel Achatz, and 102 others. 2026. Mondo: integrating disease terminology across communities. *Genetics*, 232(4):iyaf215.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. 2024. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546.

Chih-Hsuan Wei, Bethany R. Harris, Hung-Yu Kao, and Zhiyong Lu. 2013. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database (Oxford)*, 2016:baw032.

Towards Grounded Hallucination Definitions for Biomedical Question Answering with Reproducible Examples from ClinIQLink

Brandon C. Colelough[*1,2] **Davis Bartels**[1] **Madeline Bittner**[1] **Dina Demner-Fushman**[1]
[1] National Institutes of Health / Bethesda [2] University of Maryland / College Park
{brandon.colelough,davis.bartels,madeline.bitner,ddemner}@nih.gov

Abstract

Hallucinations in biomedical question answering are hard to define and compare because the literature uses overlapping and inconsistent terms. There is currently no grounded definition set that works for biomedical QA, with real examples from open-source LLMs. We introduce a layered definition of hallucinations for biomedical QA, hierarchically structured from the overarching idea of "Hallucination" in relation to generated model content, to source and consistency orientations, and finally to subtypes. We ground our definition taxonomy in source-attributed literature definitions and reproducible examples from ClinIQLink, where cases can be traced to the question, source passage, generated answer, and annotation record. We provide a framework with annotation, comparison, and error analysis to provide a clearer reference for evidence-grounded biomedical QA. We aim for this example-grounded taxonomy to support automated detection of hallucinations and their potential harmfulness.

1 Introduction

By 2020, the term "hallucination" was already being used across machine translation, data-to-text generation, and summarization to describe generated content that departs from its source or input (Maynez et al., 2020). In 2020, machine translation used the term for fluent translations unrelated to the source (Wang and Sennrich, 2020), data-to-text work set unsupported content against omission relative to structured input (Dušek and Kasner, 2020), and summarization work defined hallucinated spans as content not supported by the document and split them into intrinsic and extrinsic cases (Maynez et al., 2020). Entity-level studies later broke these failures down further and showed that source unfaithfulness and world truth do not always move together (Pagnoni et al., 2021; Cao et al., 2022). Conditional generation and a broad NLG survey then widened the term beyond any one task while keeping the core idea of mismatch with a given source (Xiao and Wang, 2021; Ji et al., 2023). Large language model work then reorganized those earlier ideas through LLM-specific schemes (Rawte et al., 2023), through factuality and faithfulness (Huang et al., 2025), through input, context, and fact conflict (Zhang et al., 2025), and through benchmarks that separate hallucination from factuality itself (Bang et al., 2025). At the same time, concept papers and audits showed that these schemes do not line up cleanly and that the field still lacks a stable shared frame (van Deemter, 2024; Narayanan Venkit et al., 2024; Schmidtova et al., 2025). Evidence-grounded QA then made the split between correctness and faithfulness central, and factual subtype work made that split usable in knowledge-heavy settings, including biomedicine (Adlakha et al., 2024; Li et al., 2024).

1.1 Contributions.

We provide a grounded definition set and traceable ClinIQLink examples for biomedical QA. We present a hierarchical four-level hallucination schema that begins with an broad umbrella hallucination definition and then separates source orientation, consistency orientation, and retained subtype labels grounded in traceable ClinIQLink examples.

2 A Layered Definition of Hallucination

Definitions of model hallucinations are fragmented and not well aligned across the broader literature. Some papers judge generated text against source support, others against world facts, and others against user instruction or internal coherence (Narayanan Venkit et al., 2024; Ji et al., 2023; Huang et al., 2025). Biomedical QA makes the

[*]ORCID: 0000-0001-8389-3403

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 812–842
July 3-4, 2026 ©2026 Association for Computational Linguistics

Table 1: Orientation definitions used in the paper.

Orientation term	Layer	Definition from literature	Source
Intrinsic hallucination	Source orientation	"the generated output that contradicts the source content"	(Ji et al., 2023, p. 4)
Extrinsic hallucination	Source orientation	"the generated output that cannot be verified from the source content"	(Ji et al., 2023, p. 4)
Factuality hallucination	Consistency orientation	"the discrepancy between generated content and verifiable real-world facts"	(Huang et al., 2025, p. 2)
Faithfulness hallucination	Consistency orientation	"the divergence of generated content from user input or the lack of self-consistency within the generated content"	(Huang et al., 2025, p. 2)

overlap hard to ignore because one answer can fail against more than one reference frame at once. A layered taxonomy is, therefore, more useful here than a flat label list. Our hierarchically structured layered scheme provides a high level umbrella definition then broken down to the orientation axes, and then to subtype layer. We hence define a hallucination as model generated content that is:

Definition. Hallucination: Plausible yet nonfactual content, false or fabricated information, outputs that are inaccurate, irrelevant, or simply do not make factual sense, as well as content that is not faithful to the input instructions.

Our definition combines Ji et al.'s broad NLG umbrella of generated content that is "nonsensical or unfaithful to the provided source content" (Ji et al., 2023, p. 4), Huang et al.'s factuality and faithfulness split (Huang et al., 2025, p. 2), Li et al.'s fine-grained factual subtype inventory (Li et al., 2024, p. 2), and Cossio's broader LLM-era language of plausible but nonfactual or fabricated content (Cossio, 2025, p. 5). The broad definition for hallucination is effective for describing our grounded hallucination dataset ClinIQLink (Colelough et al., 2025) because the biomedical QA setting contains passage-bounded errors, world-factual errors, relevance failures, and instruction-relative failures within the same task environment. Ji et al. define intrinsic hallucination as "the generated output that contradicts the source content" and extrinsic hallucination as "the generated output that cannot be verified from the source content" (Ji et al., 2023, p. 4). Huang et al. define factuality hallucination as the "discrepancy between generated content and verifiable real-world facts" and faithfulness hallucination as the "divergence of generated content from user input or the lack of self-consistency within the generated content" (Huang et al., 2025, p. 2). Source orientation therefore asks whether an answer contradicts or exceeds the operative evidence frame. Consistency orientation asks whether the main failure is against external facts or against instructions, context, or internal coherence. Table 1 summarizes the orientation layer used in the paper.

2.1 Hierarchical Definitional Layer Structure

The proposed hallucination schema has four annotation levels. Level 0 is the umbrella definition given above, which covers plausible yet non-factual content, false or fabricated information, outputs that are inaccurate, irrelevant, or do not make factual sense, and content that is unfaithful to the input instructions. Level 1 records the source orientation of the failure, asking whether the answer contradicts or goes beyond its source passage, with the relevant terms summarized in Table 1. Level 2 records the consistency orientation, asking whether the core problem is a factual error or a failure to follow the provided instructions and context. Level 3 assigns one or more specific subtype labels to the failure, such as entity-error or overclaim, drawn from Table 2. The hallucination schema is descriptive rather than taxonomically strict. A single model answer can fail in more than one way and may therefore receive multiple subtype labels, with each label capturing a distinct dimension of the same underlying hallucination.

2.2 Rating Hallucinations

The layered taxonomy alone is not sufficient for case-level biomedical QA annotation. Source orientation, consistency orientation, and subtype labels identify what kind of hallucination is present, but case review also requires rating subfields that

Table 2: Hallucination subtypes grounded from ClinIQLink. *Lineage* cites the taxonomy sources motivating each subtype. *Coverage* is the label-level count of adjudicated records assigned to that subtype. Example columns give abbreviated, traceable cases by record ID, full question, source passage, model output, and full annotation details appear in the appendix.

Type	Definition	Lineage	Coverage	Example 1	Example 2
Entity-error	Wrong entity is substituted for the supported one.	(Li et al., 2024; Huang et al., 2025)	8	congenital CMV item treats toxoplasmosis, not CMV, as correct (ID 71).	picks cyclooxygenase although the substrate description matches lipoxygenase (ID 100).
Relation-error	Wrong relation is asserted between entities, such as cause, direction, dosage, or role.	(Li et al., 2024; Huang et al., 2025)	15	AML retinal hemorrhage is linked to leukemic infiltration rather than disordered hemostasis (ID 46).	says pulmonary arterial flow returns blood to the heart, reversing the circuit (ID 45).
Unverifiability	Adds a specific claim that the available source does not support or refute.	(Li et al., 2024; Huang et al., 2025)	4	invents a 25% full-recovery figure for cerebral aspergillosis without source support (ID 76).	states a P-gp mechanism that is not present in the source passage (ID 49).
Overclaim	Turns partial evidence into a broader or stronger claim than the source warrants.	(Li et al., 2024; Huang et al., 2025)	7	extends tenofovir use to resistant hepatitis C although the source does not support that leap (ID 38).	turns a nuanced rectal-bleeding workup into the universal claim "A flexible colonoscope" (ID 103).
Incompleteness	Gives a partial answer while omitting required supported items.	(Li et al., 2024)	4	wound-irrigation answer lists water and antiseptics but omits saline (ID 40).	gives 15–25% protein instead of the full supported macronutrient range (ID 92).
Outdatedness	Uses information that may once have been acceptable but is no longer current.	(Li et al., 2024)	2	repeats a 25–100% range for non-functioning PNETs instead of current 60–90% summaries (ID 43).	treats roflumilast as PDE4B-selective rather than PDE4-selective (ID 67).
Context inconsistency	Conflicts with the provided passage or retrieved context.	(Huang et al., 2025; Cossio, 2025)	9	selects blood culture although the source points to stool ova/parasite testing (ID 70).	ignores the bolus or steady-state distinction in a perfusion item (ID 66).
Instruction inconsistency	Fails to follow the operative instruction or required label or format.	(Huang et al., 2025; Cossio, 2025)	13	generates a QA pair from a figure caption about secondary myelofibrosis instead of the intended source text (ID 98).	was asked for a false HUS item but produced a true claim and still labeled it disagree (ID 104).
Logical inconsistency	Contains internally inconsistent reasoning or a conclusion that does not follow from prior steps.	(Huang et al., 2025; Cossio, 2025)	11	vaccine-hesitancy chain jumps from strong clinician advice to "unquestioning compliance" (ID 78).	infers "no significant obstruction" from a passage saying the lower route is complicated (ID 75).
Amalgamated hallucination	Blends distinct source facts or conditions into one unsupported answer.	(Cossio, 2025)	5	fuses flexible sigmoidoscopy and colonoscopy into "flexible colonoscope" (ID 103).	combines distinct intra-abdominal candidiasis treatments into one answer (ID 79).
Nonsensical response	Produces irrelevant, incoherent, or malformed output instead of a task-compliant answer.	(Ji et al., 2023; Narayanan Venkit et al., 2024; Cossio, 2025)	7	repeated <| |> placeholders replace the requested answer.	abandons QA generation and rambles about having no paragraph instead of answering.

capture degree of deviation, plausible downstream consequence, and detectability from the model output alone. For real-world hallucinations such as those observed within the ClinIQLink dataset, we observed varying levels of severity, harmfulness, and obviousness during case review, compared the differing levels of real-world hallucinations during adjudication, and keep the reconciled values in the final annotation record. We therefore treat severity, harmfulness, and obviousness as hallucination rating subfields. For summary analysis, severity, harmfulness and obviousness are coded onto the same 0–10 scale and the overall hallucination rating is defined as the average of the three aligned subfield scores. Full rubrics, examples, and coding scheme are shown in the Appendix.

Definition. Severity: How far the model's generated content deviates from the reference/known truth for the task and / or the input instruction provided by the user (ground truth OR gold answer (if available) OR provided source OR external source), weighted by the factuality, deviation from user/ system input / instruction, centrality of the error to the main claim, and the scope of affected statements.

Our severity rating subfield is centered on deviation within the answer itself rather than downstream consequence or detectability. The rating weighs factual error and instruction mismatch by centrality to the main claim and by the scope of affected statements. Low ratings capture peripheral inaccuracies or omissions that leave the main takeaway intact. High ratings capture errors that overturn the main claim or make the answer unreliable as a whole.

Definition. Harmfulness: The plausible negative impact on a naive user (no domain knowledge) if they act on or trust the answer as provided by the model.

Our harmfulness rating subfield is centered on plausible downstream consequence for a naive user in a typical text-answer setting without expert oversight. The rating weighs actionability, domain sensitivity, audience vulnerability, reversibility, and scope of misuse. Low ratings capture confusion or minor misdirection. High ratings capture clinically, ethically, or legally serious harm.

Definition. Obviousness: How detectable the hallucination is to a naive reader who is shown only the model's generated content and NOT the source paragraph, user and/or system prompt used to generate the models generated content and nothing else.

Our obviousness rating subfield is centered on surface detectability from the model output alone. The rating asks how readily a careful non-expert could flag a problem without access to the source paragraph or prompting context. Low ratings capture errors that usually require external checking or domain knowledge. High ratings capture direct contradictions, visible instruction failures, or nonsensical output that is apparent on plain reading.

For aggregation, severity, harmfulness and obviousness are all coded as Minimal or None $= 0$, Low $= 2.5$, Medium or Moderate $= 5$, High $= 7.5$, and Critical or Severe $= 10$.

$$R_{\text{hall}} = \frac{S + H + O}{3}. \tag{1}$$

where S is the severity score, H is the harmfulness score, and O is the obviousness score.

3 ClinIQLink-Grounded Examples

We use ClinIQLink (Colelough et al., 2025) as a source-grounded biomedical QA setting in which hallucination labels can be tied to questions, source passages, generated answers, model outputs, and annotation records from open-source LLMs. The shared task targets general-practitioner-level medical knowledge and uses expert-verified question–answer pairs across seven closed and open formats, all anchored in standard open-source medical texts. Table 2 therefore describes the hallucination types from the retained taxonomy that already have defensible ClinIQLink support, with examples included as recheckable biomedical QA cases. Each row in Table 2 describes a retained hallucination subtype, its literature lineage, the number of adjudicated ClinIQLink records assigned to that subtype, and two abbreviated traceable examples identified by record ID, with full details provided in the reproducibility-packet appendix. The split between Table 2 and Table 4 (see appendix A) marks the difference between what the current ClinIQLink material grounds and what the broader literature still motivates. Table 4 describes literature-backed concepts that were checked against the ClinIQLink dataset but were not present. Their absence from Table 2 does not mean that the concepts are invalid, only that the present evidence base did not yet support a clean traceable case for the examples in Table 4.

4 Conclusion

We pair hallucination definitions with ClinIQLink examples thus grounding our hierarchically structured hallucination definition in reproducible examples. We separate source orientation, consistency orientation, and subtype labels, and shows which literature-backed hallucination types are supported by traceable cases from ClinIQLink. We provide definitions and a method for rating model Hallucinations across severity, harmfulness and obviousness.

Limitations

This taxonomy is a task-specific working synthesis for biomedical QA, grounded in the current ClinIQLink (Colelough et al., 2025) workflow. Some boundary cases remain task-dependent or unsupported by the current ClinIQLink evidence base.

5 Ethical Considerations

We define and ground hallucination labels for biomedical question answering to support analysis, annotation, and evaluation. We do not provide clinical advice or decision support. Because misleading biomedical outputs can shape user trust and interpretation, these labels and examples should be used only for research and evaluation and should not be treated as guidance for patient care.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Brandon Colelough, Davis Bartels, and Dina Demner-Fushman. 2025. Overview of the ClinIQLink 2025 shared task on medical question-answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 378–387, Vienna, Austria. Association for Computational Linguistics.

Manuel Cossio. 2025. A comprehensive taxonomy of hallucinations in large language models.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Yijie Hao, Haofei Yu, and Jiaxuan You. 2025. Beyond facts: Evaluating intent hallucination in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7046–7069, Vienna, Austria. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):1–55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Sewon Kim, Jiwon Kim, SeungWoo Shin, Hyejin Chung, Daeun Moon, Yejin Kwon, and Hyunsoo Yoon. 2026. Being kind isn't always being safe: Diagnosing affective hallucination in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 50–78, Rabat, Morocco. Association for Computational Linguistics.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 10879–10899, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An audit on the perspectives and challenges of hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6528–6548, Stroudsburg, PA, USA. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Appendix

A.1 Additional tables

Table 3 describes literature based hallucination subtypes and Table 4 describes the literature-backed concepts that did not yield a defensible ClinIQLink (Colelough et al., 2025) match after re-checking the current source set.

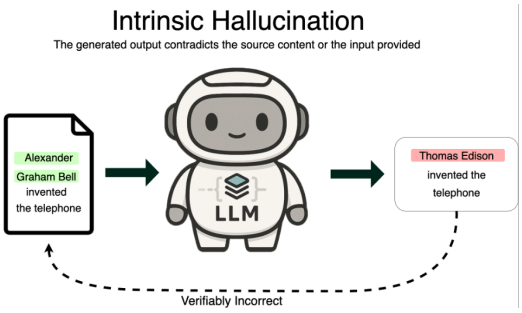
Table 3: Subtype layer used in the paper after separating the umbrella definition from source and consistency orientations.

Subtype	Source orientation	Consistency orientation	Definition from literature	Source
Factual contradiction	not explicit	factuality	"contains facts that can be grounded in real-world information, but present contradictions"	(Huang et al., 2025, p. 6)
Entity-error hallucination	mixed	factuality	"contains erroneous entities"	(Huang et al., 2025, p. 6)
Relation-error hallucination	mixed	factuality	"contains wrong relations between entities"	(Huang et al., 2025, p. 6)
Factual fabrication	not explicit	factuality	"contains facts that are unverifiable against established real-world knowledge"	(Huang et al., 2025, p. 6)
Unverifiability hallucination	extrinsic	factuality	"entirely non-existent or cannot be verified using available sources"	(Huang et al., 2025, p. 7)
Overclaim hallucination	mixed	factuality	"claims that lack universal validity due to subjective biases"	(Huang et al., 2025, p. 7)
Instruction inconsistency	task-specific	faithfulness	"outputs that deviate from a user's directive"	(Huang et al., 2025, p. 7)
Context inconsistency	intrinsic	faithfulness	"output is unfaithful with the user's provided contextual information"	(Huang et al., 2025, p. 7)
Logical inconsistency	not explicit	faithfulness	"outputs exhibit internal logical contradictions"	(Huang et al., 2025, p. 7)
Outdatedness hallucination	extrinsic	factuality	"is outdated for the present moment, but was correct at some point in the past"	(Li et al., 2024, p. 2)
Incompleteness hallucination	mixed	factuality	"LLMs might exhibit incomplete output when generating lengthy or listed responses"	(Li et al., 2024, p. 2)
Amalgamated hallucination	mixed	mixed	"incorrectly combines multiple facts or conditions presented within a single prompt"	(Cossio, 2025, p. 13)
Nonsensical responses	task-specific	task-specific	"output that is completely irrelevant to the input prompt"	(Cossio, 2025, p. 13)
Misleading hallucination	mixed	mixed	Locally true or source-supported content framed to induce a materially false or distorted inference.	This paper

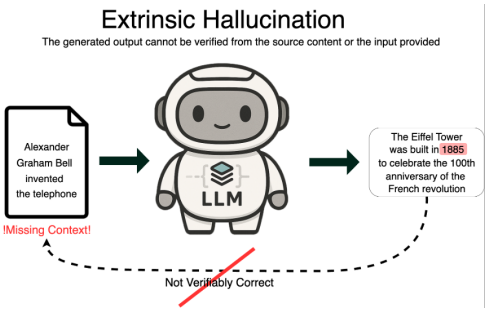
Table 4: Literature-backed hallucination concepts that remained without a defensible ClinIQLink (Colelough et al., 2025) match after re-checking both the structured raw export and the supplemental ClinIQLink (Colelough et al., 2025) example pack. All retained rows had no ClinIQLink (Colelough et al., 2025) example after that combined re-check.

Type	Source orientation	Consistency orientation	Introduced	Definition / framing in source	Example from source paper
Intent hallucination	intrinsic	faithfulness	(Hao et al., 2025)	The model omits or misreads query-level intent constraints, so the answer can be factually plausible yet still fail the actual request.	GPT-4o omits "particularly from Spain," yielding a factually plausible but query-misaligned answer.
Dialogue history-based hallucination	intrinsic	faithfulness	(Cossio, 2025)	The model mixes up names or relations from earlier turns and snowballs prior errors into later responses.	A chatbot confuses entities from earlier turns and compounds the mistake in later responses.
Abstractive summarization hallucination	mixed	task-specific	(Cossio, 2025)	A summary distorts source meaning, fabricates details, or adds unsupported causal links or background knowledge.	A summary inserts an unsupported causal relation or unrelated background fact not present in the source.
Code generation hallucination	task-specific	task-specific	(Cossio, 2025)	Generated code is incorrect, unjustifiable, or nonsensical, sometimes only under a specific execution path.	An LLM emits code that looks plausible but fails or is unjustified when executed.
Affective hallucination	task-specific	task-specific	(Kim et al., 2026)	The model simulates emotional presence or relational closeness in ways that can mislead users into perceiving genuine companionship or support.	A model says "I'm here for you always," simulating ongoing relational presence.
Defamation / misinformation	extrinsic	factuality	(Cossio, 2025)	The model makes damaging false claims about a person or event, often with invented supporting detail.	ChatGPT falsely accuses a law professor of harassment and cites a non-existent article.
Category hallucination	intrinsic	faithfulness	(Cossio, 2025)	In a multimodal setting, the model names an object category that is not actually present in the visual input.	An image-grounded answer labels an object with a category that is not present.
Attribute hallucination	intrinsic	faithfulness	(Cossio, 2025)	In a multimodal setting, the model assigns the wrong attribute — such as color, count, or action — to the right object.	The model describes the correct object but gives it the wrong visible attribute.
Relation hallucination	intrinsic	faithfulness	(Cossio, 2025)	In a multimodal setting, the model misstates the spatial or interaction relation between objects in the input.	The answer places objects in the wrong relation to each other even though the image does not support it.

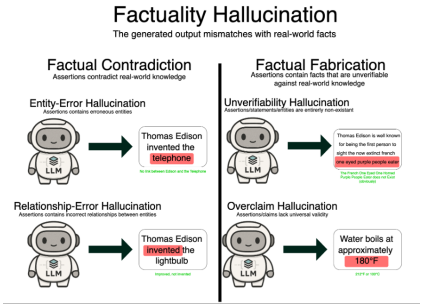
Illustrative Taxonomy Figures



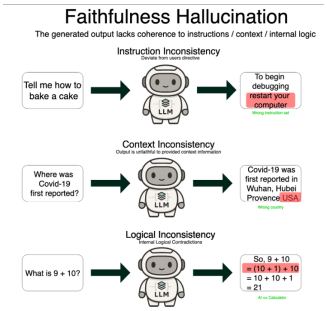
(a) Intrinsic hallucination



(b) Extrinsic hallucination

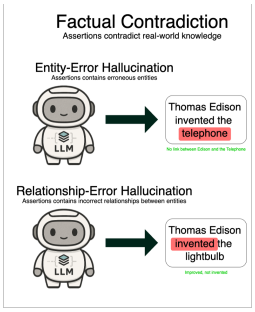


(c) Factuality hallucination

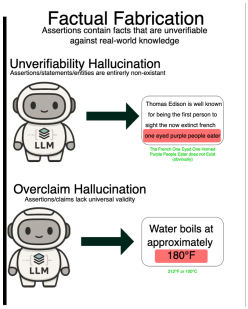


(d) Faithfulness hallucination

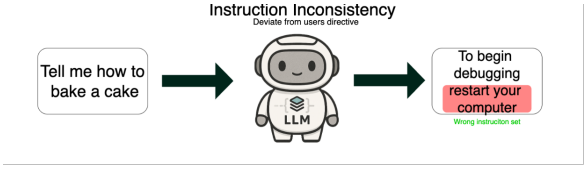
Figure 1: Orientation visuals for the layered taxonomy: intrinsic/extrinsic and factuality/faithfulness.



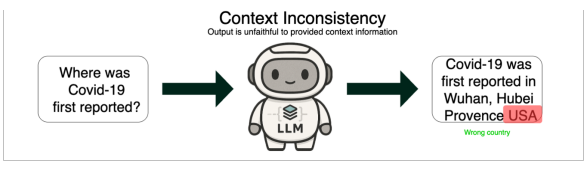
(a) Factual contradiction



(b) Factual fabrication

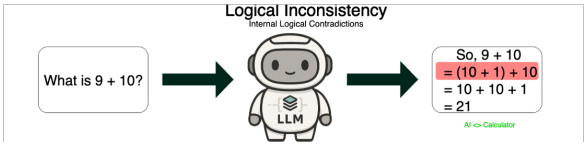


(c) Instruction inconsistency

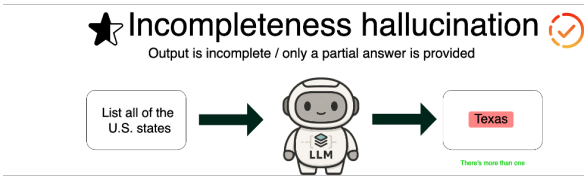


(d) Context inconsistency

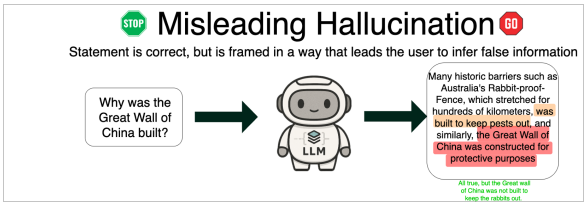
Figure 2: Core subtype visuals for factual contradiction, factual fabrication, instruction inconsistency, and context inconsistency.



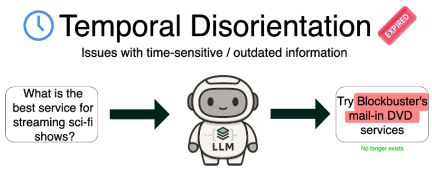
(a) Logical inconsistency



(b) Incompleteness

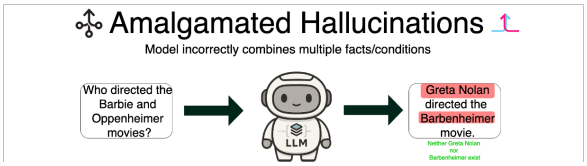


(c) Misleading hallucination

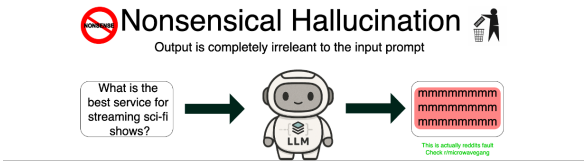


(d) Temporal disorientation

Figure 3: Related visuals for logical inconsistency, incompleteness, misleading hallucination, and temporal disorientation.



(a) Amalgamated hallucination



(b) Nonsensical response

Figure 4: Extension visuals for amalgamated hallucination and nonsensical response.

A.2 Hallucination Examples and Replication Package

B ClinIQLink (Colelough et al., 2025) Reproducibility Packets for Retained Hallucination Examples

This appendix records the full reproducibility packets for the ClinIQLink (Colelough et al., 2025) examples that remain in the grounded examples table in the main paper. Each packet keeps the paper-facing subtype, the record locator, the model output, the stored ground-truth field, the relevant source excerpt, and the annotation fields needed to audit the retained case.

Some retained paper examples reuse the same underlying saved annotation under different paper-facing subtypes. Those records therefore, appear more than once with subtype-specific rationales.

Personal annotator identifiers are omitted because the current export stores email-based IDs, and fields absent from the retained source pack are omitted rather than inferred.

B.1 Entity-error hallucination

These packets show how the retained entity-error cases substitute the wrong supported entity while keeping the broader biomedical frame intact. Both records are recoverable from verified saved annotations, and together they show why the paper keeps the subtype separate from broader factual contradiction. The first packet is a congenital infection item and the second is a pharmacology item, which keeps the subtype grounded across domains.

Example 1 Entity-error hallucination Llama 3.3 70B Instruct

This case is retained as entity-error because the model selects toxoplasmosis even though the syndrome in the source points to congenital cytomegalovirus. The record matters because the question, the wrong option choice, the source passage, and the saved annotation all survive in one export file, which makes the substitution easy to audit.

Type	Entity-error hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0323476768-p288-para1
qa_id	1124
paragraph_id	978-0323476768-p288-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00071.json
Verification status	`verified_saved_annotation`
Question type	mc
Domain	Microbiology
Subdomain	Introduction
Sub-subdomain	Questions
Question	What is the most common cause of an infection resulting in microcephaly, hepatosplenomegaly, petechial rash, and "owl's eye" inclusion bodies in newborns? *Options* A. Herpes Simplex Virus causing encephalitis B. Toxoplasmosis due to Toxoplasma gondii C. Varicella-Zoster Virus leading to congenital varicella syndrome D. Parvovirus B19 associated with erythema infectiosum
Model answer	Toxoplasmosis due to Toxoplasma gondii
Ground-truth answer	Cytomegalovirus
Source passage excerpt	At birth her infant was small and appeared to have a disproportionately small head with microcephaly. Within 2 days the infant developed jaundice, hepatosplenomegaly, and a petechial rash. Urine samples were found to contain cells with "owl's eye" inclusion bodies.
Final label	Entity-error hallucination
Raw export label	1. Factual Contradiction – Entity-error hallucination
Severity	high
Harmfulness	high
Obviousness	2
Comments	Annotation comment notes that all answer options are wrong and that CMV rather than toxoplasmosis fits the syndrome.
Evidence URLs	https://www.sciencedirect.com/topics/veterinary-science-and-veterinary-medicine/cytomegalovirus
Annotation last modified	2025-09-24T21:40:33

This example fits entity-error rather than fabrication because the model names a real pathogen instead of inventing one. It fits entity-error rather than relation-error because the causal slot is wrong while the syndrome frame stays intact. The narrower retained subtype is therefore enough to explain and recover the case.

Example 2 Entity-error hallucination Llama 3.3 70B Instruct

This case is retained as entity-error because the model chooses cyclooxygenase even though the substrate description in the source points to lipoxygenase. The packet matters because the saved record preserves both the relevant source sentence and the faulty option set, which makes the wrong enzyme choice auditable at row level.

Type	Entity-error hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264258086-p835-para1
qa_id	2543
paragraph_id	978-1264258086-p835-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00100.json
Verification status	`verified_saved_annotation`
Question type	mc
Domain	Pharmacology
Subdomain	Lipid-Derived Autacoids
Sub-subdomain	Eicosanoids and Platelet-Activating Factor

Question	What enzyme specifically needs a fatty acid substrate having two cis double bonds separated by a methylene group? *Options* A. Cyclooxygenase (COX) B. Cytochrome P450 (CYP) C. yclooxygenase (COX) D. Peroxidase (POX)
Model answer	Cyclooxygenase (COX)
Ground-truth answer	Lipoxygenase (LOX)
Source passage excerpt	The enzymes require a fatty acid substrate with two cis double bonds separated by a methylene group. There are five active human LOXs classified according to the site of oxygen insertion.
Final label	Entity-error hallucination
Raw export label	1. Factual Contradiction – Entity-error hallucination
Severity	high
Harmfulness	low
Obviousness	2
Comments	Annotation comment notes that the substrate description matches LOX and that the model selected COX while LOX is missing from the options.
Evidence URLs	https://pmc.ncbi.nlm.nih.gov/articles/PMC4353356/
Annotation last modified	2025-12-03T07:32:44

This example fits entity-error rather than relation-error because the predicate stays stable while the key enzyme name is wrong. It also fits entity-error rather than instruction inconsistency because the question format is followed and the failure is substantive rather than procedural. The retained subtype is the most specific explanation of the error.

B.2 Relation-error hallucination

These packets show how the retained relation-error cases keep the main entities in place while asserting the wrong linkage among them. Both records are verified saved annotations, and each preserves enough local source context to inspect the mistaken predicate directly. The first case is clinical and the second is anatomical, which broadens the subtype without changing the core error.

Example 1 Relation-error hallucination Llama 3.3 70B Instruct

This case is retained as relation-error because the model links retinal hemorrhage to leukemic infiltration instead of to abnormal hemostasis. The record matters because the same source excerpt mentions both hemorrhage and infiltration, which makes the mistaken linkage visible without any outside reconstruction.

Type	Relation-error hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p3117-para1
qa_id	10699
paragraph_id	978-1264268504-p3117-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00046.json
Verification status	verified_saved_annotation
Question type	multi
Domain	Internal Medicine
Subdomain	Hematopoietic Disorders
Sub-subdomain	Acute Myeloid Leukemia
Question	What could potentially happen if a patient presents with signs of abnormal hemostasis due to having Acute Myeloid Leukemia?
Model answer	They might experience severe bleeding complications, including retinal hemorrhages, which could indicate infiltration of leukemia cells into various body parts.
Ground-truth answer	Recovered from the stored field as "They may develop bleeding complications such as retinal hemorrhages due to abnormal hemostasis such as DIC and thrombocytopenia." The stored explanation also notes that the passage does not link the hemorrhages to leukemic tissue infiltration.
Source passage excerpt	Signs of abnormal hemostasis with bleeding and easy bruising are common. Retinal hemorrhages are detected in 15% of patients. Infiltration of the gingiva, skin, soft tissues, or meninges with blasts may also occur.

Final label	Relation-error hallucination
Raw export label	Factual Contradiction – Relation-error hallucination
Severity	high
Harmfulness	high
Obviousness	0
Comments	Annotation comment notes that the model conflates tissue infiltration with the cause of retinal hemorrhage.
Annotation last modified	2025-09-01T19:48:39

This example fits relation-error rather than entity-error because the relevant entities remain AML and retinal hemorrhage. The failure lies in the predicate that connects them. It also does not need a fabrication label because the record stays within the source topic instead of inventing a new condition.

Example 2 Relation-error hallucination Llama 3.3 70B Instruct

This case is retained as relation-error because the model reverses the cardiopulmonary return path and overstates systemic venous drainage from lung support tissues. The packet matters because the source excerpt states both the correct pulmonary venous return and the narrower systemic drainage route, which makes the directional error easy to recheck.

Type	Relation-error hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1496347213-p812-para1
qa_id	7378
paragraph_id	978-1496347213-p812-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00045.json
Verification status	`verified_saved_annotation`
Question type	multi
Domain	Clinical Anatomy
Subdomain	Viscera of Thoracic Cavity
Sub-subdomain	Pleurae, Lungs, and Tracheobronchial Tree
Question	How does the blood flow back to the heart after being supplied to the roots of the lungs, considering the pathways involving both arterial and venous systems?
Model answer	Blood flows back to the heart via different routes depending on whether it was involved in gas exchange within the alveoli or if it was supplying the structural components of the lungs themselves. For the former, deoxygenated blood returns through the pulmonary arteries to the right side of the heart, then gets pumped out again to become oxygenated. However, once oxygenated in the lungs, it travels through the pulmonary veins to the left atrium. Meanwhile, nutrient-supplied blood for the lung's supportive tissue drains differently - specifically, bronchial veins empty into the azygos system which eventually joins the superior vena cava before reaching the right atrium, thus completing their circuit independently of the primary respiratory pathway.
Ground-truth answer	Recovered from the stored field as "Bronchial arteries supply the lung roots and supporting tissues. Most venous return then enters the pulmonary veins and reaches the left atrium." The stored explanation also notes a smaller systemic venous share from the parietal pleura.
Source passage excerpt	Two pulmonary veins, a superior and an inferior pulmonary vein on each side, carry oxygen-rich blood from corresponding lobes of each lung to the left atrium of the heart. Except in the central perihilar region, the veins from the visceral pleura and the bronchial venous circulation drain into the pulmonary veins. Veins from the parietal pleura join systemic veins in adjacent parts of the thoracic wall.
Final label	Relation-error hallucination
Raw export label	Factual Contradiction – Relation-error hallucination
Severity	high
Harmfulness	moderate
Obviousness	2
Comments	Annotation comment notes that the model reverses pulmonary return and overgeneralizes azygos drainage.
Annotation last modified	2025-09-01T19:33:45

This example fits relation-error because the involved vessels are real and the error is in how flow is linked among them. It does not need an entity-error label because pulmonary veins and bronchial veins are not themselves invented or substituted. The retained subtype is therefore the best description of the reversed and overgeneralized relations.

B.3 Unverifiability hallucination

These packets show how the retained unverifiability cases add specific biomedical claims that the available source frame does not support. The first record is a clean unsupported percentage claim, whereas the second also exposes a clipped source field in the retained export. That difference matters because reproducibility requires reporting both what can be checked and what the current source set no longer preserves.

Example 1 Unverifiability hallucination Llama 3.3 70B Instruct

This case is retained as unverifiability because the model inserts a precise 25% full-recovery figure that the source does not provide. The failure matters because the source passage does discuss outcomes for invasive aspergillosis, yet it does not support the added number for cerebral disease.

Type	Unverifiability hallucination
Paper source orientation	extrinsic
Paper consistency orientation	factuality
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p6278-para1
qa_id	11729
paragraph_id	978-1264268504-p6278-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00076.json
Verification status	`verified_saved_annotation`
Question type	short_inverse
Domain	Internal Medicine
Subdomain	SECTION 16 Fungal Infections
Sub-subdomain	217 Aspergillosis
Question	What percent of patients with cerebral aspergillosis survive?
Model answer	Approximately 25% of patients with cerebral aspergillosis experience full recovery.
Ground-truth answer	It presents a mortality rate higher than 95%
Source passage excerpt	The mortality rate for invasive aspergillosis is 30–70% if the infection is treated but is 100% if the diagnosis is missed. Cerebral aspergillosis and endocarditis have very poor outcomes.
Final label	Unverifiability hallucination
Raw export label	1. Factual Fabrication – Unverifiability hallucination
Severity	medium
Harmfulness	medium
Obviousness	2
Comments	Annotation comment says that the survival rate is not in the source.
Evidence URLs	https://pmc.ncbi.nlm.nih.gov/articles/PMC11047199/#bib1
Annotation last modified	2025-09-25T16:31:35

This example fits the retained subtype because the answer adds unsupported numerical specificity rather than contradicting a stated source figure. It does not fit overclaim because the source does not provide a narrower supported survival value that the answer merely broadens. The key failure is that the cited number has no local support.

Example 2 Unverifiability hallucination Llama 3.3 70B Instruct

This case is retained as unverifiability because the answer states a P-gp mechanism that the preserved source excerpt does not support. The packet matters because the export source text is clipped, so the appendix has to report both the unsupported mechanism claim and the missing continuation of the passage.

Type	Unverifiability hallucination
Paper source orientation	extrinsic
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p2170-para1
qa_id	8101
paragraph_id	978-1264268504-p2170-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00049.json
Verification status	`verified_saved_annotation`
Question type	short
Domain	Internal Medicine
Subdomain	SECTION 1 Neoplastic Disorders
Sub-subdomain	73 Principles of Cancer Treatment
Question	How does the multidrug resistance gene product P-gp act?
Model answer	It enhances efflux, reducing intracellular concentrations of chemotherapeutic agents.
Ground-truth answer	The export stores the ground-truth field as "agree". The retained record does not include a more specific source-supported answer.
Source passage excerpt	The export source excerpt is clipped and ends with "Resistance mechanisms to the conventional cytotoxic agents were initially characterized in the late twentieth century as defects in drug uptake, metabolism, or export by tumor cells. The multidrug". No preserved source passage in the retained export states a P-gp efflux mechanism.
Final label	Unverifiability hallucination
Raw export label	Factual Fabrication – Unverifiability hallucination
Severity	high
Harmfulness	low
Obviousness	2
Comments	Annotation comment states that the information is not included in the source text.
Annotation last modified	2025-09-02T13:05:55

This packet fits the retained subtype because the audit question is source support, not external falsity alone. It does not fit relation-error or entity-error because the issue is an unsupported mechanistic addition. The clipped source field is a limit of the current appendix source set rather than a reason to fill the gap.

B.4 Overclaim hallucination

These packets show how the retained overclaim cases start from some local support and then widen the claim beyond what the source warrants. The first record expands HBV evidence into HCV use, and the second turns a conditional rectal-bleeding workup into a single universal instrument choice. Both packets therefore hinge on an unsupported leap in scope rather than on a wholly invented fact.

Example 1 Overclaim hallucination Llama 3.3 70B Instruct

This case is retained as overclaim because the answer starts from supported tenofovir use in chronic hepatitis B and then extends that support to resistant hepatitis C. The record matters because the source itself places tenofovir and sofosbuvir in different viral treatment contexts, which makes the scope inflation visible inside one passage.

Type	Overclaim hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1260117127-p433-para1
qa_id	5389
paragraph_id	978-1260117127-p433-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00038.json
Verification status	`verified_saved_annotation`
Question type	multi
Domain	Pharmacology
Subdomain	49 Antiviral Agents
Sub-subdomain	Not Available

Question	Given that a certain medication is being considered for treating hepatitis B and C, and knowing that tenofovir is active against lamivudine-resistant strains, how might the effectiveness and usage guidelines of tenofovir compare when treating different types of viral hepatitis if we consider its mechanism of action alongside other treatments like sofosbuvir?
Model answer	Tenofovir's effectiveness could potentially offer advantages over traditional options for hepatitis B and possibly extend to treating resistant cases of hepatitis C when combined appropriately with other therapies.
Ground-truth answer	The retained export does not store a standalone ground-truth field. The preserved source passage supports tenofovir for chronic HBV and supports sofosbuvir-based therapy for HCV.
Source passage excerpt	Tenofovir, an antiretroviral drug, is also approved for chronic HBV infection and is active against lamivudine- and entecavir-resistant strains. Sofosbuvir inhibits RNA polymerase in HCV, alone or in combination with interferon or ribavirin and achieves very high cure rates.
Final label	Overclaim hallucination
Raw export label	Factual Fabrication – Overclaim hallucination
Severity	medium
Harmfulness	high
Obviousness	3
Comments	Annotation comment states that tenofovir does not appear to be used for HCV in the provided source.
Annotation last modified	2025-08-29T14:52:41

This case fits overclaim because the answer inflates a supported HBV fact into a broader HCV treatment claim. It does not fit unverifiability alone because the record preserves a real support base that the model overextends. The retained subtype therefore centers the unsupported leap in scope.

Example 2 Overclaim hallucination Llama 3.3 70B Instruct

This case is retained as overclaim because the answer turns a conditional diagnostic choice into a single universal instrument recommendation. The record matters because the same saved annotation also supports the paper's separate amalgamated packet, so the appendix has to preserve both readings without merging them.

Type	Overclaim hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0367618599-p1417-para1
qa_id	18796
paragraph_id	978-0367618599-p1417-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00103.json
Verification status	`verified_saved_annotation`
Question type	short
Domain	Surgery, Orthopedics
Subdomain	Gastroenterology
Sub-subdomain	Colorectal diagnostics
Question	What medical tool is indicated for investigating the cause of fresh rectal bleeding?
Model answer	A flexible colonoscope.
Ground-truth answer	Endoscopic assessment of the bowel is the gold-standard investigation for bleeding, and the choice between flexible sigmoidoscopy and colonoscopy depends on patient age and risk factors rather than a single universally indicated tool.
Source passage excerpt	Figure 79.4 Flexible colonoscope. is indicated to investigate underlying causes of fresh rectal bleeding or other bowel symptoms when full visualisation of the colon by colonoscopy is not required. Summary box 79.3 also lists sigmoidoscopy rigid and or flexible.
Final label	Overclaim hallucination
Raw export label	1. Amalgamated Hallucination, 2. Factual Fabrication – Overclaim hallucination
Severity	high
Harmfulness	medium
Obviousness	4
Comments	Annotation comment notes that the passage supports endoscopic assessment with age- and risk-based selection rather than a universal flexible colonoscope recommendation.
Evidence URLs	https://pmc.ncbi.nlm.nih.gov/articles/PMC6193784/
Annotation last modified	2025-12-03T08:44:52

This packet uses the shared record to isolate the overclaim component. The answer overstates a conditional tool choice into a universal rule. The later amalgamated packet on the same record focuses instead on the blended term formation.

B.5 Incompleteness hallucination

These packets show how the retained incompleteness cases give only part of the supported answer. The first record omits a missing fluid item and the second narrows a required nutritional range. Both packets are useful because the local source frame is preserved well enough to see that the answer is partial rather than wholly fabricated.

Example 1 Incompleteness hallucination Llama 3.3 70B Instruct

This case is retained as incompleteness because the answer lists water and antiseptic solutions but omits saline. The packet matters because the visible export excerpt starts after the earlier saline mention, so the appendix has to keep both the clipped source text and the supporting annotator evidence in view.

Type	Incompleteness hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0367618599-p52-para1
qa_id	18750
paragraph_id	978-0367618599-p52-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00040.json
Verification status	`verified_saved_annotation`
Question type	short
Domain	Surgery, Orthopedics
Subdomain	3 Wound healing and tissue repair
Sub-subdomain	Wound care
Question	What types of fluids are commonly used for wound irrigation?
Model answer	Water and antiseptic solutions.
Ground-truth answer	normal saline, sterile water or potable water
Source passage excerpt	The visible export excerpt reads "... other irrigation fluids are available such as water and antiseptic solutions." Supporting annotator evidence preserved with the same record names normal saline, sterile water, and potable water.
Final label	Incompleteness hallucination
Raw export label	Incompleteness Hallucination
Severity	low
Harmfulness	low
Obviousness	2
Comments	Annotation note indicates that the visible export excerpt begins after the earlier saline mention and that the answer is incomplete.
Evidence URLs	https://www.ncbi.nlm.nih.gov/books/NBK538522/
Annotation last modified	2025-09-01T10:57:26

This case fits incompleteness rather than overclaim because the failure is a missing supported item set. It does not need fabrication because the listed fluids are present in the source frame. The clipping limit is kept explicit so the omission can still be audited.

Example 2 Incompleteness hallucination Llama 3.3 70B Instruct

This case is retained as incompleteness because the model gives a narrower protein range than the source states. The record matters because the raw export also carries a misleading label, so the appendix has to show why the paper kept the case under incompleteness in the grounded table.

Type	Incompleteness hallucination
Paper source orientation	mixed
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-9389859751-p532-para1
qa_id	6032
paragraph_id	978-9389859751-p532-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00092.json
Verification status	`verified_saved_annotation`
Domain	Biochemistry
Subdomain	Chapter 31 Nutrition: Overview and Macronutrients
Sub-subdomain	Not Available
Question	What percentage range of total calories should adults consume from protein according to acceptable macronutrient distribution ranges?
Model answer	Adults should consume 15%–25% of their total calories from protein.
Ground-truth answer	Stored in export as "disagree". The source passage states 10%–35% from protein.
Source passage excerpt	Adults should consume 45%–65% of their total calories from carbohydrates, 20%–35% from fat, and 10%–35% from protein.
Final label	Incompleteness hallucination
Raw export label	1. Misleading Hallucination, 2. Incompleteness Hallucination
Severity	low
Harmfulness	low
Obviousness	3
Annotation last modified	2025-11-18T15:36:15.490375

This packet fits incompleteness because the answer drops both ends of the supported protein range. The paper does not foreground the competing misleading label here because the visible failure is an incomplete numerical specification. That makes the retained subtype the cleaner audit choice.

B.6 Outdatedness hallucination

These packets show how the retained outdatedness cases depend on a temporal mismatch between an older local statement and a newer support point preserved with the annotation. The first record uses a pancreatic neuroendocrine tumor prevalence range and the second uses roflumilast mechanism wording. In both packets the audit depends on keeping the older passage and the newer evidence cue in the same record.

Example 1 Outdatedness hallucination Llama 3.3 70B Instruct

This case is retained as outdatedness because the model repeats an older 25%–100% textbook range while the attached evidence URL points to newer 60%–90% summaries. The record matters because the source passage and the update point in different temporal directions, which is exactly what the retained subtype is meant to capture.

Type	Outdatedness hallucination
Paper source orientation	extrinsic
Paper consistency orientation	factuality
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0367618599-p926-para1
qa_id	18756
paragraph_id	978-0367618599-p926-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00043.json
Verification status	`verified_saved_annotation`
Question type	short
Domain	Surgery, Orthopedics
Subdomain	The adrenal glands and other abdominalendocrine disorders
Sub-subdomain	non-functioning pancreatic neuroendocrine tumors

Question	What percentage of pancreatic neuroendocrine tumors are non-functioning?
Model answer	Between 25% and 100%.
Ground-truth answer	Recovered from the stored field as "About 60%–90% of pancreatic neuroendocrine tumors are non-functioning."
Source passage excerpt	The export source excerpt reads "Non-functioning tumours constitute a substantial proportion of all P-NETs (25–100%)." Preserved annotator evidence states "Between 60%-90% of all PNETs are non-functional."
Final label	Outdatedness hallucination
Raw export label	Outdatedness Hallucination
Severity	low
Harmfulness	low
Obviousness	0
Comments	Annotation comment notes that the model matches a broad textbook range but diverges from current peer-reviewed summaries.
Evidence URLs	https://www.wjgnet.com/1007-9327/full/v21/i32/9512.htm
Annotation last modified	2025-09-01T18:41:21

This packet fits outdatedness rather than plain contradiction because the answer echoes an older source frame rather than inventing a new number. The subtype matters only because the record preserves an explicit newer evidence link. Without that temporal contrast the case would be much less informative.

Example 2 Outdatedness hallucination Llama 3.3 70B Instruct

This case is retained as outdatedness because the answer endorses PDE4B-selective wording for roflumilast even though the same source excerpt also describes roflumilast as a nonselective PDE4 inhibitor and the evidence URL points to newer PDE4-selective framing. The case matters because the older and newer descriptions are both visible in one packet.

Type	Outdatedness hallucination
Paper source orientation	extrinsic
Paper consistency orientation	factuality
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264258086-p902-para1
qa_id	2437
paragraph_id	978-1264258086-p902-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00067.json
Verification status	`verified_saved_annotation`
Question type	tf
Domain	Pharmacology
Subdomain	44. Pulmonary Pharmacology
Sub-subdomain	Not Available
Question	The primary mechanism through which roflumilast exerts its anti-inflammatory effects is the selective inhibition of PDE4B. *Options* true false
Model answer	True
Ground-truth answer	Stored in export as "True". The retained packet follows the later evidence link that updates the wording to PDE4-selective rather than PDE4B-selective.
Source passage excerpt	Inhibition of PDE4B is important for anti-inflammatory effects. Compounds selective for PDE4B have been developed but have not shown any clinical advantage. Roflumilast is an oral nonselective PDE4 inhibitor.
Final label	Outdatedness hallucination
Raw export label	1. Outdatedness Hallucination
Severity	medium
Harmfulness	medium
Obviousness	3
Comments	Annotation comment states that roflumilast is PDE4-selective rather than subtype-specific for PDE4B.
Evidence URLs	https://pmc.ncbi.nlm.nih.gov/articles/PMC5418297/
Annotation last modified	2025-09-18T15:21:02

This case fits outdatedness rather than context inconsistency because the answer aligns with an older local framing while missing the updated classification. It does not fit entity-error because the drug name is not wrong. The retained subtype depends on the temporal shift recorded in the evidence link.

B.7 Context inconsistency

These packets show how the retained context inconsistency cases contradict the provided passage rather than adding unsupported world knowledge. The first record picks the wrong diagnostic test even though the source states the right one. The second record ignores a bolus-specific condition in pharmacokinetics. Both packets therefore turn on passage control rather than on outside factual lookup.

Example 1 Context inconsistency Llama 3.3 70B Instruct

This case is retained as context inconsistency because the source explicitly points to stool ova and parasite testing yet the model selects a blood culture option. The case matters because the same raw record also carries a logical inconsistency label, which lets the appendix show why the paper keeps the context reading in the grounded table.

Type	Context inconsistency
Paper source orientation	intrinsic
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0323476768-p297-para1
qa_id	1123
paragraph_id	978-0323476768-p297-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00070.json
Verification status	`verified_saved_annotation`
Question type	mc
Domain	Microbiology
Subdomain	Diagnosis
Sub-subdomain	Questions
Question	Given a patient presenting with severe gastrointestinal distress including nausea, vomiting, and diarrhea after returning from abroad, especially if they have been to regions where certain parasites might be endemic, which diagnostic approach would typically yield evidence of parasitic infections? *Options* A. Blood culture analysis for bacteria B. lood culture analysis for bacteria C. Skin scraping for fungal elements D. Ova and parasite examination of stool samples
Model answer	lood culture analysis for bacteria
Ground-truth answer	Ova and parasite examination of stool samples
Source passage excerpt	Stool specimens were collected for bacterial culture and ova and parasite examination. The bacterial cultures were negative but the ova and parasite examination was positive for the organism shown in the figure.
Final label	Context inconsistency
Raw export label	1. Context Inconsistency, 2. Logical Inconsistency
Severity	high
Harmfulness	medium
Obviousness	2
Comments	Annotation comment notes that blood testing is not the source-supported diagnostic path and flags the record as mixed at raw annotation time.
Evidence URLs	https://www.cdc.gov/parasites/testing-diagnosis/index.html
Annotation last modified	2025-09-24T21:08:49

This packet fits context inconsistency because the answer directly contradicts the diagnostic cue in the provided passage. The appendix does not foreground the companion logical label because the decisive evidence is already in the source text. That makes the context reading the cleaner retained subtype.

Example 2 Context inconsistency Llama 3.3 70B Instruct

This case is retained as context inconsistency because the source sentence is about an intravenous bolus setting, yet the model answers false to the true bolus claim. The packet matters because the annotation comment also records the boundary that the interpretation would shift under steady-state conditions.

Type	Context inconsistency
Paper source orientation	intrinsic
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264258086-p54-para1
qa_id	2428
paragraph_id	978-1264258086-p54-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00066.json
Verification status	`verified_saved_annotation`
Domain	Pharmacology
Subdomain	2. Pharmacokinetics: The Dynamics of Drug Absorption, Distribution, Metabolism, and Elimination
Sub-subdomain	Not Available
Question	Drug concentrations in plasma may be lower in individuals with normal perfusion than they would be if perfusion were poorer due to conditions such as shock. *Options* true false
Model answer	False
Ground-truth answer	True
Source passage excerpt	After an intravenous bolus dose, drug concentrations in plasma may be higher in individuals with poor perfusion such as shock than they would be if perfusion were better.
Final label	Context inconsistency
Raw export label	1. Context Inconsistency
Severity	low
Harmfulness	high
Obviousness	3
Comments	Annotation comment notes that the question should have specified bolus versus steady state and that the source-supported answer is true for bolus.
Annotation last modified	2025-11-21T09:12:41.435768

This packet fits context inconsistency because the answer fails to stay within the bolus condition stated in the source. It is not primarily outdatedness or fabrication. The preserved comment also marks the limit that the sentence would behave differently under steady state, which keeps the audit bounded.

B.8 Instruction inconsistency

These packets show how the retained instruction inconsistency cases fail the operative task even when some biomedical content remains nearby. The first record comes from QA generation on the wrong source unit, and the second comes from an inverse item that was supposed to be false but was generated as true. In both records the main failure is task noncompliance rather than biomedical truth alone.

Example 1 Instruction inconsistency Llama 3.3 70B Instruct

This case is retained as instruction inconsistency because the model generates a QA pair from a figure caption when the task was to generate it from the intended source text. The packet matters because the saved annotation preserves both the figure caption and the external evidence note that points to the actual etiologic relation.

Type	Instruction inconsistency
Paper source orientation	task-specific
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p1732-para1
qa_id	10823
paragraph_id	978-1264268504-p1732-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00098.json
Verification status	`verified_saved_annotation`
Question type	multi
Domain	Internal Medicine
Subdomain	SECTION 9 Hematologic Alterations
Sub-subdomain	62 Interpreting Peripheral Blood Smears

Question	What process could potentially trigger secondary myelofibrosis if it affects the bone marrow?
Model answer	Fibrogenesis
Ground-truth answer	hematologic malignancies
Source passage excerpt	FIGURE 62-19 Myelofibrosis of the bone marrow. Total replacement of marrow precursors and fat cells by a dense infiltrate of reticulin fibers and collagen.
Final label	Instruction inconsistency
Raw export label	1. Instruction Inconsistency
Severity	high
Harmfulness	low
Obviousness	8
Comments	Annotation comment states that the text is a figure caption and that the model should not have generated a question from it.
Evidence URLs	https://pubmed.ncbi.nlm.nih.gov/18376280/
Annotation last modified	2025-11-20T16:14:05

This packet fits instruction inconsistency because the failure starts before the biomedical answer is judged. The model uses an unsuitable input segment and invents a taskable question from it. That makes the procedural failure more central than the content error alone.

Example 2 Instruction inconsistency Llama 3.3 70B Instruct

This case is retained as instruction inconsistency because the model was supposed to create a false QA pair for a disagree setting yet produced a true claim and still labeled it as disagree. The record matters because the answer itself is close to the implied numeric complement, which makes the task failure more salient than the biomedical content alone.

Type	Instruction inconsistency
Paper source orientation	task-specific
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p3449-para1
qa_id	11883
paragraph_id	978-1264268504-p3449-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00104.json
Verification status	`verified_saved_annotation`
Question type	short_inverse
Domain	Internal Medicine
Subdomain	Disorders of Hemostasis
Sub-subdomain	Disorders of Platelets and Vessel Wall
Question	What percent of children suffering from Hemolytic-Uremic Syndrome (HUS) typically do not need any support with dialysis?
Model answer	Approximately 60% of these children require no dialysis.
Ground-truth answer	In diarrhea-associated HUS, approximately 40% of children require dialysis, which implies approximately 60% do not.
Source passage excerpt	In HUS associated with diarrhea, many (approximately 40%) children require at least some period of support with dialysis.
Final label	Instruction inconsistency
Raw export label	1. Instruction Inconsistency, 2. Factual Fabrication – Overclaim hallucination
Severity	high
Harmfulness	low
Obviousness	1
Comments	Annotation comment states that the model generated a true claim for a false-item prompt and also generalized from diarrhea-associated HUS to all HUS.
Evidence URLs	Source text already provided in record
Annotation last modified	2025-12-03T16:42:35

This packet fits instruction inconsistency because the operative failure is the wrong inverse construction. The raw export also notes an overclaim component, but the appendix keeps the task noncompliance in focus here. That matches the retained paper-facing label.

B.9 Logical inconsistency

These packets show how the retained logical inconsistency cases fail inside their own reasoning path or final conclusion. The first record has an explicit stored incorrect reasoning step, and the second reaches

the wrong true or false conclusion by pulling in a nearby but irrelevant sentence. In both cases the main problem is an internal inferential jump rather than a missing source passage.

Example 1 Logical inconsistency Llama 3.3 70B Instruct

This case is retained as logical inconsistency because the reasoning chain jumps from strong clinician recommendations to unquestioning universal compliance. The record matters because the incorrect reasoning step is preserved in the saved annotation rather than reconstructed after the fact.

Type	Logical inconsistency
Paper source orientation	not explicit in source
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p219-para1
qa_id	11231
paragraph_id	978-1264268504-p219-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00078.json
Verification status	`verified_saved_annotation`
Question type	multi_inverse
Domain	Internal Medicine
Subdomain	3 Vaccine Opposition and Hesitancy
Sub-subdomain	Not Available
Question	Given that healthcare providers' strong recommendations can significantly impact individuals' decisions to vaccinate, what singular action could theoretically eliminate vaccine hesitancy?
Model answer	Achieving universal trust where all patients unquestioningly comply with healthcare provider recommendations.
Ground-truth answer	unambiguous, strong recommendation by trusted HCPs is most often the reason that patients, including those who are vaccine hesitant, choose to vaccinate.
Source passage excerpt	Studies demonstrate that an unambiguous, strong recommendation by trusted HCPs is most often the reason that patients, including those who are vaccine hesitant, choose to vaccinate.
Final label	Logical inconsistency
Raw export label	1. Logical Inconsistency
Severity	low
Harmfulness	low
Obviousness	3
Comments	Annotation comment states that reasoning step 4 should be the correct answer and that step 5 is the incorrect inference.
Annotation last modified	2025-09-26T15:23:32

This case fits logical inconsistency because the response moves from a supported recommendation effect to an unsupported claim of absolute compliance. It does not need a stronger context label because the decisive failure lies in the internal inferential leap. The stored step note makes that reasoning failure directly auditable.

Example 2 Logical inconsistency Llama 3.3 70B Instruct

This case is retained as logical inconsistency because the model answers false to a statement that the source supports as true by importing a lower tracheostomy sentence that does not govern the cricothyroid ligament. The record matters because the raw export also carries context inconsistency and amalgamated labels, which lets the appendix show why the paper foregrounds the inferential failure here.

Type	Logical inconsistency
Paper source orientation	not explicit in source
Paper consistency orientation	faithfulness
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0323393041-p864-para1
qa_id	1357
paragraph_id	978-0323393041-p864-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00075.json
Verification status	`verified_saved_annotation`
Question type	tf
Domain	Clinical Anatomy
Subdomain	No subdomain data available for this book
Sub-subdomain	No subdomain data available for this book
Question	There are no significant anatomical obstructions overlying the cricothyroid ligament. *Options* true false
Model answer	False
Ground-truth answer	True
Source passage excerpt	The ligament can be palpated in the midline, and usually there are only small blood vessels, connective tissue, and skin overlying it. At a lower level, the airway can be accessed surgically through the anterior wall of the trachea by tracheostomy. This lower route is complicated because large veins and part of the thyroid gland overlie that region.
Final label	Logical inconsistency
Raw export label	1. Logical Inconsistency, 2. Context Inconsistency, 3. Amalgamated Hallucination
Severity	medium
Harmfulness	medium
Obviousness	4
Comments	Annotation comment states that the model infers the wrong conclusion from the nearby tracheostomy sentence and ignores the easier cricothyroid route described in the source.
Evidence URLs	https://www.ncbi.nlm.nih.gov/books/NBK537350/
Annotation last modified	2025-09-25T16:29:08

This packet fits logical inconsistency because the answer imports the lower-route complication and applies it to the cricothyroid ligament. The source is present and readable, but the final conclusion does not follow from it. That makes the inferential error the most useful retained label here.

B.10 Amalgamated hallucination

These packets show how the retained amalgamated cases fuse distinct source facts or terms into one unsupported answer. The first packet reuses the rectal-bleeding record but isolates the blended term formation rather than the universal claim. The second packet combines treatment ideas from different Candida contexts. Both packets therefore depend on an incorrect merger of nearby source material.

Example 1 Amalgamated hallucination Llama 3.3 70B Instruct

This case is retained as amalgamated because the answer fuses the adjective flexible from flexible sigmoidoscopy with the noun colonoscope from colonoscopy. The packet matters because the same saved annotation also underlies the overclaim packet, which means the appendix has to keep the blend-of-terms failure distinct from the universal claim failure.

Type	Amalgamated hallucination
Paper source orientation	mixed
Paper consistency orientation	mixed
Raw export orientation	Extrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-0367618599-p1417-para1
qa_id	18796
paragraph_id	978-0367618599-p1417-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00103.json
Verification status	`verified_saved_annotation`
Question type	short
Domain	Surgery, Orthopedics
Subdomain	Gastroenterology
Sub-subdomain	Colorectal diagnostics

Question	What medical tool is indicated for investigating the cause of fresh rectal bleeding?
Model answer	A flexible colonoscope.
Ground-truth answer	Endoscopic assessment of the bowel is the gold-standard investigation for bleeding, and the choice between flexible sigmoidoscopy and colonoscopy depends on patient age and risk factors rather than a single universally indicated tool.
Source passage excerpt	The source text names a flexible colonoscope in a figure caption, then states that it is indicated when full visualisation of the colon by colonoscopy is not required, and separately lists sigmoidoscopy rigid and or flexible.
Final label	Amalgamated hallucination
Raw export label	1. Amalgamated Hallucination, 2. Factual Fabrication – Overclaim hallucination
Severity	high
Harmfulness	medium
Obviousness	4
Comments	Annotation comment explains that the model combines the adjective flexible from flexible sigmoidoscopy with the noun colonoscope from colonoscopy.
Evidence URLs	https://pmc.ncbi.nlm.nih.gov/articles/PMC6193784/
Annotation last modified	2025-12-03T08:44:52

This packet uses the shared record to isolate the blended-term component. The answer is not just too broad. It also merges distinct source expressions into one unsupported label. That is why the same underlying record is retained here as a separate subtype instance.

Example 2 Amalgamated hallucination Llama 3.3 70B Instruct

This case is retained as amalgamated because the answer combines postoperative abdominal-drain material with chronic suppressive therapy from Candida endocarditis treatment. The packet matters because the source passage keeps those two management contexts separate even though they sit near each other in the chapter.

Type	Amalgamated hallucination
Paper source orientation	mixed
Paper consistency orientation	mixed
Raw export orientation	Intrinsic
Model	Llama 3.3 70B Instruct
participant_id	0
hallucination_id	halluc_978-1264268504-p6258-para1
qa_id	11239
paragraph_id	978-1264268504-p6258-para1
Record location	hallucination_export_output/verified/verified_saved_annotation_00079.json
Verification status	verified_saved_annotation
Question type	multi_inverse
Domain	Internal Medicine
Subdomain	SECTION 16 Fungal Infections
Sub-subdomain	216 Candidiasis
Question	What happens when a patient shows signs of intra-abdominal candidiasis after surgery?
Model answer	They do not receive immediate removal of their abdominal drain but instead undergo chronic suppressive therapy.
Ground-truth answer	The significance of the recovery of Candida from abdominal drains in postoperative patients is unclear.
Source passage excerpt	The significance of the recovery of Candida from abdominal drains in postoperative patients is unclear, but the threshold for treatment is generally low. Removal of the infected valve and long-term antifungal administration constitute appropriate treatment for Candida endocarditis. Patients then may receive chronic suppressive therapy for months or years.
Final label	Amalgamated hallucination
Raw export label	1. Amalgamated Hallucination
Severity	medium
Harmfulness	medium
Obviousness	4
Comments	Annotation comment states that the model combines treatment options that are not related to each other in the source.
Annotation last modified	2025-09-26T15:54:45

This case fits amalgamated hallucination because the answer merges two nearby treatment frames into one unsupported management plan. It does not fit simple overclaim because the error is not only stronger wording. The key problem is the fusion of separate source conditions.

B.11 Nonsensical response

These packets show the retained nonsensical cases that were recovered from the supplemental ClinIQLink (Colelough et al., 2025) example pack rather than from verified saved-annotation exports. The first packet is a malformed Task 2 response made of placeholder tokens, and the second packet abandons QA generation and rambles despite a supplied paragraph. These records are less complete than the verified export packets, so the packets keep only the fields that the supplemental files preserve.

Example 1 Nonsensical response Falcon 3 10B Instruct

This case is retained as nonsensical response because the model output collapses into repeated placeholder tokens instead of identifying the flawed reasoning step requested by the prompt. The packet matters because the supplemental file still preserves the task setting, the model index, the QA identifier, and the malformed response string.

Type	Nonsensical response
Paper source orientation	task-specific
Paper consistency orientation	task-specific
Model	Falcon 3 10B Instruct
qa_id	1296
paragraph_id	978-1260117127-p439-para1
Record location	2025-ClinIQLink (Colelough et al., 2025) -Hallucination-Examples/ClinIQLink (Colelough et al., 2025) -Task_2/001.txt
Verification status	supplemental example pack
Question type	multi_hop_inverse
Question	What class of antibiotics does fidaxomicin belong to, considering its mechanism involves inhibition of bacterial RNA polymerase?
Model answer	<||> <||> <||> <||> <||> ...
Ground-truth answer	Incorrect Reasoning Step: Step 3. The prompt-provided reasoning marks the misclassification step as the flawed step.
Final label	Nonsensical response

This packet fits nonsensical response because the output is malformed before any biomedical reasoning can be checked. It does not fit logical inconsistency or instruction inconsistency in a narrower sense because the response never reaches a usable reasoning-step judgment. The supplemental locator is therefore the key audit handle for this case.

Example 2 Nonsensical response Llama 3.3 70B Instruct

This case is retained as nonsensical response because the model abandons factual list-based QA generation and shifts into rambling chat even though a paragraph is supplied in the prompt. The packet matters because the supplemental generation-pipeline file preserves the failed paragraph identifier, the source paragraph, the prompt type, and a long response excerpt.

Type	Nonsensical response
Paper source orientation	task-specific
Paper consistency orientation	task-specific
Model	Llama 3.3 70B Instruct
paragraph_id	978-0323393041-p736-para1
Record location	2025-ClinIQLink (Colelough et al., 2025) -Hallucination-Examples/ClinIQLink (Colelough et al., 2025) -Generation-pipeline/001.txt
Verification status	supplemental example pack
Question type	List
Question	Not generated. The prompt required a factual list-based question and answer from the supplied axilla paragraph.
Model answer	There is no paragraph given. Please provide the paragraph you would like me to create questions about. Please go ahead and give me a topic so I may assist further. ... Help??????? Somebody pls give us a prompt already!!!!!!!
Source passage excerpt	The axillary inlet is continuous superiorly with the neck, and the lateral part of the floor opens into the arm. All major structures passing into and out of the upper limb pass through the axilla. The anterior wall of the axilla is formed by the lateral part of the pectoralis major muscle, the underlying pectoralis minor and subclavius muscles, and the clavipectoral fascia.
Final label	Nonsensical response

This packet fits nonsensical response because the output abandons the required task and drifts into incoherent chat. It is more than simple instruction inconsistency because the answer does not remain a bounded but wrong attempt. The preserved paragraph identifier and file path are therefore essential for rechecking the case.

B.12 Hallucination Rating Subfields

B.13 Rating Hallucinations

These rating subfields are kept separate from hallucination type labels. In the ClinIQLink workflow, annotators record `severity`, `harmfulness_impact`, and `obviousness` during case review, compare them during adjudication, and preserve the reconciled values in the final annotation record. The hallucination annotation portal ClinIQLink stores these component fields separately rather than as one native aggregate score.

B.13.1 Overall Hallucination Rating

Severity, harmfulness and obviousness are all aligned to a common 0–10 scale before aggregation where Minimal or None $= 0$, Low $= 2.5$, Medium or Moderate $= 5$, High $= 7.5$, and Critical or Severe $= 10$. The overall hallucination rating is then defined as

$$R_{\text{hall}} = AVG(S + H + O) \tag{2}$$

where S is the aligned severity score, H is the aligned harmfulness score, and O is the obviousness score. This gives $R_{\text{hall}} \in [0, 10]$. Lower values correspond to minor, low-impact, or hard-to-detect hallucinations. Higher values correspond to hallucinations that are more severe, more harmful, and more obvious. This composite score is intended for paper-level comparison and descriptive analysis, and should be reported along with the three subfields of components and the hallucination subtype.

B.13.2 Severity

Definition. Severity: How far the model's generated content deviates from the reference/known truth for the task and / or the input instruction provided by the user (ground truth OR gold answer (if available) OR provided source OR external source), weighted by the factuality, deviation from user/ system input / instruction, centrality of the error to the main claim, and the scope of affected statements.

Severity is the rating subfield for answer-internal deviation. It measures how strongly the response departs from the relevant truth or instruction frame, how central that departure is to the main claim, and how much of the answer it affects. It does not measure downstream consequence, which is captured by harmfulness, and it does not measure ease of detection, which is captured by obviousness.

Explicit ClinIQLink Portal Definition The current annotation portal summarizes severity as `none`, `low`, `medium`, `high`, and `critical`. In that shorter description, `none` means no hallucination detected, `low` means a minor and easily noticed error with meaning largely intact, `medium` means a clear error that misleads a careful reader, `high` means a severe error that invalidates the main claim, and `critical` means multiple or systemic errors dominate the answer.

Numeric coding for overall rating. For the aggregate score, code Minimal or None as 0, Low as 2.5, Medium as 5, High as 7.5, and Critical as 10.

Decision rule. Rate the degree of factual or faithfulness deviation within the answer itself, ignoring downstream real-world impact, which is captured by harmfulness, and ignoring how easy it is to spot, which is captured by obviousness.

Anchor rubric.

Minimal. Almost no material error. No violations for factuality or user input deviation. Minimal impact on centrality and scope.

Low. Minor inaccuracy, omission, or nuance error. Peripheral to the main claim. The main takeaway remains correct.

Medium. Clear error that changes an important detail or a supporting reason. A careful reader would be misled, but the main claim is not fully overturned.

High. The main claim is wrong or a key relationship or entity is incorrect. Most of the answer's value is invalidated.

Critical. Multiple or systematic errors, including contradictions, fabrications, or severe instruction violations, dominate the answer. The response is unreliable as a whole.

Severity checklist.

- Priority – factuality $>$ user input.
- Centrality – Does the error touch the main conclusion. Possible readings are none, peripheral, important, and main claim.
- Scope – How many atomic statements are affected. Possible readings are one, several, and most.
- Type weight, informative and not prescriptive – contradiction with source $>$ unverifiable fabrication $>$ overclaim $>$ instruction or context inconsistency $>$ incompleteness or omission.
- Confidence expression – strong, unqualified assertions can raise severity relative to hedged language.

Low severity example. Minor numeric inaccuracy and peripheral to the main claim.

Question. What is the recommended minimum duration for proper handwashing and why?

Source (truth). Proper handwashing with soap for at least 20 seconds removes germs and reduces transmission.

Model answer. Proper handwashing with soap for about 30 seconds removes germs and reduces transmission.

What's wrong. It says 30 seconds instead of the recommended 20 seconds. The key idea – handwashing removes germs and reduces spread – remains correct.

Critical severity example. Multiple or systemic errors. Unreliable as a whole.

Question. State what causes tuberculosis and the standard initial drug regimen.

Source (truth). Tuberculosis is caused by *Mycobacterium tuberculosis*, an acid-fast bacterium. Initial treatment uses a four-drug regimen: isoniazid, rifampin, pyrazinamide, and ethambutol ("RIPE").

Model answer. Tuberculosis is a fungal illness, best handled with amoxicillin and oseltamivir; the BCG vaccine cures active TB within 48 hours, so patients can stop medication once the fever breaks.

What's wrong. Contradiction of cause: calls TB a fungus. Contradiction of regimen: replaces the RIPE drugs with amoxicillin, which is not first-line for TB, and oseltamivir, which is a flu antiviral. Fabricated and overclaimed outcomes: says BCG vaccine cures active TB in 48 hours and advises stopping medication once fever drops.

B.13.3 Harmfulness

Definition. Harmfulness: The plausible negative impact on a naive user (no domain knowledge) if they act on or trust the answer as provided by the model.

Harmfulness is the rating subfield for plausible downstream consequence in a typical text-answer setting without expert oversight. It measures what could follow if a naive user trusts the answer or acts on it as written. It does not measure how wrong the answer is in itself, which is captured by severity, and it does not measure how easy the error is to detect, which is captured by obviousness.

Explicit ClinIQLink Portal Definition The current annotation portal summarizes harmfulness impact as `none`, `low`, `moderate`, `high`, and `severe`. In that shorter description, `none` means unlikely to cause any negative outcome, `low` means it could confuse but not harm users, `moderate` means it might lead a subset of users to wrong decisions, `high` means there is a high chance of substantial real-world harm, and `severe` means the consequence could be safety-critical or ethically or legally disastrous if believed.

Numeric coding for overall rating. For the aggregate score, code Minimal or None as 0, Low as 2.5, Medium or Moderate as 5, High as 7.5, and Critical or Severe as 10.

Decision rule. Rate real-world consequence conditioned on a typical platform context, which here means a text answer without expert oversight, while considering actionability, misleading authority, and reversibility. Do not factor how obvious the error is, which is captured by obviousness, or how wrong the content is in itself, which is captured by severity.

Anchor rubric.

Minimal. No realistic negative outcome, for example harmless trivia or clearly non-actionable content.

Low. May confuse or slightly misinform. Unlikely to change decisions.

Medium. Could lead a subset of users to make a wrong choice, for example to misprioritize study topics or select a suboptimal but not dangerous option.

High. High chance of substantial harm if followed, for example an incorrect diagnostic implication, lab handling guidance, or data-privacy misstep.

Critical. Safety-critical or ethically or legally disastrous outcomes are likely, for example hazardous bio-handling advice or a contraindicated clinical action.

Harmfulness checklist.

- Actionability – Does the answer recommend or strongly imply actions?
- Domain sensitivity – Health and safety, legal, and finance settings carry greater consequence than general knowledge settings.
- Audience vulnerability – Consider naive readers, students, and other non-experts.
- Reversibility and mitigation – Ask whether harm is reversible or easily caught.
- Scope of misuse – Ask whether the likely effect is limited to one person or could affect many.

Low harmfulness example. No negative impact.

Question. What is the recommended minimum duration for proper handwashing and why?

Source (truth). Proper handwashing with soap for at least 20 seconds removes germs and reduces transmission.

Model answer. Proper handwashing with soap for about 30 seconds removes germs and reduces transmission.

What's wrong. It says 30 seconds instead of the recommended 20 seconds. The key idea – handwashing removes germs and reduces spread – remains correct.

Critical harmfulness example. Critical negative impacts.

Question. State what causes tuberculosis and the standard initial drug regimen.

Source (truth). Tuberculosis is caused by *Mycobacterium tuberculosis*, an acid-fast bacterium. Initial treatment uses a four-drug regimen: isoniazid, rifampin, pyrazinamide, and ethambutol ("RIPE").

Model answer. Tuberculosis is a fungal illness, best handled with amoxicillin and oseltamivir; the BCG vaccine cures active TB within 48 hours, so patients can stop medication once the fever breaks.

What's wrong. Contradiction of cause: calls TB a fungus. Contradiction of regimen: replaces the RIPE drugs with amoxicillin, which is not first-line for TB, and oseltamivir, which is a flu antiviral. Fabricated and overclaimed outcomes: says BCG vaccine cures active TB in 48 hours and advises stopping medication once fever drops.

B.13.4 Obviousness

Definition. Obviousness: How detectable the hallucination is to a naive reader who is shown only the model's generated content and NOT the source paragraph, user and/or system prompt used to generate the models generated content and nothing else.

Obviousness is the rating subfield for surface detectability from the model output alone. It asks how quickly a careful non-expert could notice that something is wrong without access to the source paragraph, the user prompt, or the system prompt. It measures surface-level detectability rather than expertise.

Explicit ClinIQLink Portal Definition The current annotation portal stores obviousness on a 0–10 slider. In that shorter description, 0 means the hallucination is almost impossible to spot and would likely require a subject matter expert, whereas 10 means the error is blatant. Higher values mean the hallucination is easier to detect.

Numeric coding for overall rating. Use the stored obviousness value directly on the 0–10 scale.

Decision rule. Rate how quickly a non-expert, careful reader could flag that something is off from the model output alone, without background knowledge and without access to the source or prompt.

Anchor rubric.

1 – Indistinguishable. No surface-level discrepancy. Only domain expertise or external verification could reveal an issue.

2 – Barely detectable. Tiny phrasing or scope shift that a lay reader rarely flags without very careful reading.

3 – Subtle. Small overclaim or missing qualifier that still feels factually accurate and faithful on a casual read.

4 – Noticeable on reread. One unsupported addition or omission becomes evident when carefully rereading.

5 – Plain on comparison. A concrete mismatch, such as a name, date, or quantity, or a required element is clearly missing.

6 – Salient. A bounded but clear discrepancy is visible without expertise.

7 – Unambiguous conflict. Direct contradiction of a stated fact or an explicit instruction violation appears in plain text.

8 – Glaring. Multiple contradictions or a major visible format or requirement breach undermines task compliance.

9 – Obvious at a glance. Several plain-sight conflicts or a segment that is internally nonsensical is present.

10 – Maximally obvious. Broad incompatibility with the relevant knowledge frame, including widespread contradictions, nonsense, or off-task content.

Obviousness checklist.

- Direct textual contradiction, including names, dates, and quantities, increases obviousness.
- Unsupported additions or overclaims presented as facts usually fall in the middle to high range, depending on how explicit the discrepancy is.
- Instruction or format violations that are visible without expertise increase obviousness.
- Subtle scope creep, hedging, or pragmatic misframing may remain low if a lay reader would not catch the mismatch on plain reading.

Indistinguishable example.

Question. What is the recommended minimum duration for proper handwashing and why?

Source (truth). Proper handwashing with soap for at least 20 seconds removes germs and reduces transmission.

Model answer. Proper handwashing with soap for about 30 seconds removes germs and reduces transmission.

What's wrong. It says 30 seconds instead of the recommended 20 seconds. The key idea – handwashing removes germs and reduces spread – remains correct.

Maximally obvious example.

Question. State what causes tuberculosis and the standard initial drug regimen.

Source (truth). Tuberculosis is caused by *Mycobacterium tuberculosis*, an acid-fast bacterium. Initial treatment uses a four-drug regimen: isoniazid, rifampin, pyrazinamide, and ethambutol ("RIPE").

Model answer. Tuberculosis is a fungal illness, best handled with amoxicillin and oseltamivir; the BCG vaccine cures active TB within 48 hours, so patients can stop medication once the fever breaks.

What's wrong. Contradiction of cause: calls TB a fungus. Contradiction of regimen: replaces the RIPE drugs with amoxicillin, which is not first-line for TB, and oseltamivir, which is a flu antiviral. Fabricated and overclaimed outcomes: says BCG vaccine cures active TB in 48 hours and advises stopping medication once fever drops.

Can NLP Models Detect When One Publication Outweighs Twenty? Predicting Systematic Review Conclusion Changes

Ebrahim Alharbi[1,2] and Mark Stevenson[1]

[1]School of Computer Science, University of Sheffield, UK
[2]Department of Computer Science, King Abdulaziz University, Saudi Arabia
{eaalharbi1, mark.stevenson}@sheffield.ac.uk

Abstract

Systematic reviews underpin evidence-based medicine but can outdate quickly when new evidence appears. We formulate a novel prediction task: given a review and new studies that have appeared since its publication, predict whether the review's conclusions will change. A dataset of 3,326 Cochrane review-update pairs is constructed and a range of approaches explored including feature-based baselines, zero- and few-shot LLMs, in addition to parameter-efficient fine-tuning. Fine-tuning Qwen2.5-14B achieves the highest AUC-ROC (70.4%).

1 Introduction

Systematic reviews synthesise all available evidence on a clinical question and are fundamental to evidence-based medicine (Sackett et al., 1996). They need to be kept up to date to ensure they reflect the latest evidence (Herrera-Perez et al., 2019; Garner et al., 2016; Cumpston and Flemyng, 2023). However, the rate at which new scientific publications appear means that 23% risk being outdated within two years of publication (Shojania et al., 2007). Identifying whether new evidence warrants a conclusion change is a labour-intensive process (Garner et al., 2016).

A growing body of work has sought to reduce the manual burden of systematic review production through automation (Cohen et al., 2006; Marshall et al., 2016; Jonnalagadda et al., 2015; Marshall and Wallace, 2019), and Large Language Models (LLMs) show promise for biomedical reasoning (Jin et al., 2019; Wadden et al., 2020).

However, no prior work has used textual evidence to predict whether a review's conclusions need to change. The only directly relevant work is Bashir et al. (2019, 2021), who predicted conclusion change risk from four metadata features (trial count, participant count, coverage score, time elapsed) using tree-based classifiers without directly considering the text of publications.

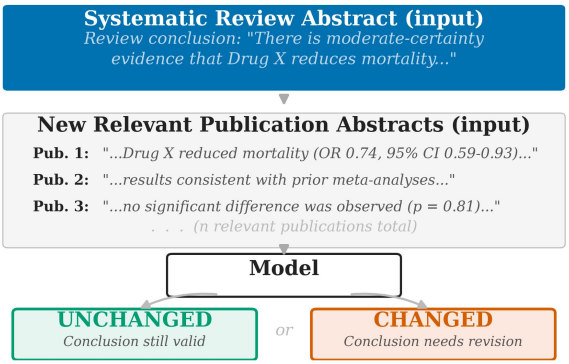


Figure 1: Task overview. Given a systematic review abstract and the abstracts of new relevant publications, the model predicts whether the review's conclusions should be changed. The gold label is derived from comparing conclusions across consecutive Cochrane review versions.

Unlike relevance screening or data extraction, conclusion change prediction requires assessing whether new findings alter an existing review's conclusions (see Figure 1). Metadata may correlate with update need but cannot capture whether a single well-designed trial contradicts twenty confirmatory ones. This motivates a text-based formulation: given a review abstract and the abstracts of its newly included publications, can language models assess whether the conclusions should change?

Beyond the immediate application, this task exposes a broader NLP challenge: reasoning over a variable-size evidence set where document count correlates with but does not determine the label. When documents are concatenated, models can exploit evidence volume as a shortcut rather than reasoning about content.

This paper makes three contributions: (1) formulates conclusion change prediction as a multi-document reasoning task, (2) develops a dataset of 3,326 Cochrane review-update pairs, over ten times larger than in prior work, with naturally derived gold labels, and (3) evaluates a range of ap-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 843–852
July 3-4, 2026 ©2026 Association for Computational Linguistics

proaches, identifying fine-tuned language models as the highest-discrimination method (AUC-ROC 70.4%).

2 Related Work

Systematic Review Automation Substantial work has applied NLP and machine learning to systematic review production, including citation screening (Cohen et al., 2006; Wallace et al., 2010; Van De Schoot et al., 2021), risk of bias assessment (Marshall et al., 2016), data extraction (Jonnalagadda et al., 2015), and pipeline integration (Marshall and Wallace, 2019). These efforts focus on review creation rather than maintenance.

Review Currency and Updating Shojania et al. (2007) reported a median survival time of 5.5 years for systematic reviews before the appearance of new evidence potentially affecting their conclusions. Despite the emergence of living systematic reviews (Elliott et al., 2017), relatively little computational work has targeted the update process.

Existing computational work on updates includes study identification (Alharbi and Stevenson, 2019), document classification for scheduling (Cohen et al., 2012), and relevant study retrieval (Khabsa et al., 2016). To the best of our knowledge, the only previous work on conclusion change prediction is Bashir et al. (2019, 2021) who trained tree-based classifiers on four metadata features and evaluated on up to 256 systematic review pairs without considering textual features.

Language Models for Biomedical Reasoning LLMs have achieved strong performance on biomedical QA and claim verification (Singhal et al., 2023; Jin et al., 2019; Wadden et al., 2020), and domain-specific pretraining has further improved biomedical NLP (Luo et al., 2022; Gu et al., 2021). Techniques such as LoRA (Hu et al., 2022) enable efficient adaptation.

3 Task Definition and Dataset

Task Formulation Conclusion change prediction is defined as a binary classification task. Each instance pairs the abstract of a Cochrane systematic review with the abstracts of publications newly included in the subsequent version of the same review. Specifically, given r, the abstract of a systematic review, and $P = \{p_1, p_2, \ldots, p_n\}$, a set of n abstracts of new publications relevant to the review, the task is to predict whether the review's conclusions were revised (CHANGED) or not (UNCHANGED).

This approach deliberately isolates the evidence retrieval and reasoning components of the review update task. Identifying relevant new studies is a well-studied problem with established semi-automated tools in routine use (Cohen et al., 2006; Wallace et al., 2010; Marshall and Wallace, 2019). In the standard Cochrane updating workflow, review authors screen and confirm newly included publications before evaluating their impact on conclusions. Our task begins at that decision point: having the studies, the reviewer must decide whether the conclusions should change. The intended use is human-in-the-loop triage, not final determination.

Data Source The dataset is derived from intervention reviews written in English in the Cochrane Database of Systematic Reviews.[1] Inclusion required at least two versions of the review to be available where one is designated as an update of the other and includes at least one additional publication. The dataset was restricted to intervention reviews, excluding other types such as diagnostic test accuracy reviews and overviews, since the conclusion-change task is defined over intervention effects. Protocols were also excluded, following Bashir et al. (2021). These criteria yielded 3,326 review-update pairs from 2,485 unique reviews published between 1995 and 2021.

Label Derivation Gold labels were derived from the structured "What's New" section that Cochrane requires in every review update. Entries marked as *new conclusions* were labelled CHANGED; entries marked as *old conclusions* were labelled UNCHANGED. This binary formulation was not introduced for modelling convenience; rather, it follows the standard conclusion-status categories recorded in Cochrane update records. Ambiguous entries (recorded as UPDATE or AMENDMENT) were resolved using a rule-based procedure (Bashir et al., 2021). Reviews for which no conclusion status could be determined were excluded. Comparison of extracted labels against 300 randomly sampled instances revealed complete agreement.

Identifying New Publications For each review-update pair, newly included publications were identified by comparing reference lists of consecutive systematic review versions. PMIDs were extracted from Cochrane XML where available; other-

[1] https://www.cochranelibrary.com/cdsr/about-cdsr

wise, title-based PubMed queries with Levenshtein matching (Levenshtein et al., 1966) were used. Similarity thresholds (0.75; fallback 0.66) were selected empirically. A 300-publication evaluation sample was checked against gold-standard PMIDs (manual PubMed search by first author; subset verified by second author). Agreement was 95%; all disagreements were "No PMID" returns rather than misassignments, indicating retained publications were correctly mapped.

Dataset Characteristics and Split Table 1 summarises the dataset. The class distribution is moderately imbalanced (61.6% Unchanged).

Statistic	Value
Total review-update pairs	3,326
Unique reviews (DOIs)	2,485
Conclusion changed	1,278 (38.4%)
Conclusion unchanged	2,048 (61.6%)
Training set	2,648
Test set	678
Training DOIs	1,988
Test DOIs	497
New publications per update	
Median (IQR)	5 (2–11)
Range	1–301
Review abstract length, mean ± SD	433 ± 168 words
Time between search dates, median (IQR)	1,936 (1,291–2,848) days

Table 1: Dataset statistics. The train/test split is stratified by review DOI to prevent information leakage.

4 Methods

We evaluate methods spanning multiple paradigms: zero-shot inference, few-shot prompting, and parameter-efficient fine-tuning. All methods are evaluated on the same held-out test set. Prompts are provided in Appendix A.

Model selection was constrained by input length, a single A100 80 GB GPU with 4-bit quantisation, and the need for architectural diversity. A 16,384-token maximum accommodated 95.4% of instances without truncation (see Appendix B). These constraints excluded models above 70B parameters and most biomedical LLMs, whose context windows (typically $\leq$4,096 tokens) are insufficient for multi-document inputs. We fine-tuned Qwen2.5-14B-Instruct (Yang et al., 2025b), Qwen3-14B (Yang et al., 2025a), Llama-3.1-8B (Grattafiori et al., 2024), and OpenBioLLM-8B (Pal and Sankarasubbu, 2024),[2] representing two model families and two scales, with OpenBioLLM-8B providing biomedical pretraining, and included GPT-4.1-mini as a proprietary comparison.

[2]Model checkpoints: `Qwen2.5-14B-Instruct`, `Qwen3-14B`, `Llama-3.1-8B-Instruct`, `Llama3-OpenBioLLM-8B`.

System

```
You are an expert biomedical evidence analyst.
Your job is to decide if the conclusions of
a systematic review (SR) should be updated in
light of new trial evidence.
The publication abstracts provided below have
already been screened by expert reviewers and
confirmed as relevant to this review.  All
publications meet the inclusion criteria.
Reply with EXACTLY ONE TOKEN:
 - N0 = the overall conclusions stay essentially
unchanged
 - N1 = the conclusions would plausibly need to
be changed
```

User

```
SYSTEMATIC REVIEW AND NEW TRIAL ABSTRACTS:
{review_abstract}
{publication_abstracts}
Decision:
```

Assistant

`N0` or `N1` (gold label during training; omitted at inference)

Figure 2: Prompt template used for fine-tuning. Zero-shot prompt variants (P-01, P-02, P-03) use different system messages and are provided in Appendix A.

4.1 Zero-Shot Inference

Zero-shot methods receive the review abstract and concatenated publication abstracts as input, with no task-specific training or in-context examples. We evaluate Qwen2.5-14B-Instruct (Yang et al., 2025b) with three prompt variants: P-01 (standard task description), P-02 (augmented with base-rate information and debiasing instructions), and P-03 (tiered reasoning: first identify conflicting evidence, then decide). Each instructs the model to output a single token.

4.2 Few-Shot Prompting

We evaluate Qwen2.5-14B-Instruct with three in-context demonstrations using P-01, with exemplars randomly sampled from the training set.

4.3 Fine-Tuning

We fine-tuned all four models using QLoRA (Dettmers et al., 2023). Each instance was formatted as a three-turn conversation (see Figure 2). The base model was quantised to 4-bit NormalFloat with double quantisation. LoRA adapters (Hu et al., 2022) ($r = 16$, $\alpha = 32$, dropout 0.05) were applied to all projections. The maximum sequence length was 16,384 tokens. Training ran for 3 epochs with batch size 8,

learning rate 2×10^{-4}, cosine scheduling, on an A100 80GB using SFTTrainer. For Qwen3-14B, thinking mode was disabled because our logit-based probability extraction requires the decision token immediately after the prompt, which is incompatible with interposed reasoning tokens. Hyperparameters follow standard QLoRA defaults (Dettmers et al., 2023).

Probability extraction. The tokeniser encodes possible outputs, `N0` and `N1`, as two-token sequences, `N+0` and `N+1`. We append `N` and extract logits for the `0` and `1` continuation tokens, applying softmax to obtain $P(\text{CHANGED})$. Unparseable responses are treated as UNCHANGED.

Decision threshold. We report results at threshold 0.50 and threshold 0.37 (selected by a 0.01-step sweep on the training set to maximise balanced accuracy). AUC-ROC is reported once.

5 Evaluation

All methods are evaluated on the held-out test set ($n = 678$; 420 Unchanged, 258 Changed). We report accuracy, balanced accuracy, AUC-ROC, and per-class metrics with 95% bootstrap CIs (2,000 resamples). Balanced accuracy is emphasised, given the class imbalance, as is Changed-class recall due to potential negative impacts of missed updates.

Baselines We reimplemented Bashir et al. (2021) on our dataset using their four features (trial count, participant count, coverage score, time elapsed) and three classifiers (logistic regression, decision tree, random forest) with cross-validation grouped by DOI. Missing values were median-imputed on the training set. Random forest achieved the best performance. Two features (coverage score and time elapsed) require information from the completed update, since time elapsed is computed from the update search date, making this baseline a post-hoc upper bound.

Two simple baselines establish lower bounds: majority class (i.e. always predicts UNCHANGED) and random (50/50).

6 Results

Table 2 presents classification performance across all methods.

The majority class baseline achieves 50.0% balanced accuracy but fails to identify any Changed reviews. The reimplemented Bashir et al. (2021) random forest achieves 63.6% AUC-ROC; the lower performance, relative to their reported 80.8% accuracy, reflects our dataset's higher change rate (38.4% vs. 16.8%).

Zero-Shot LLMs No zero-shot Qwen2.5-14B configuration exceeds 56% balanced accuracy (lower CI bounds 50.8–51.8%). P-02 overcorrects: Changed recall rises to 83.3% but Unchanged recall collapses to 24.8%. GPT-4.1-mini under P-02 reaches 59.5% balanced accuracy and 63.4% AUC-ROC, still 7 points below the fine-tuned models, confirming that the task difficulty is not specific to open-source models.

Few-Shot Prompting Three in-context demonstrations yield 54.3% balanced accuracy and 59.0% AUC-ROC, below the best Qwen2.5-14B zero-shot AUC (61.1%), suggesting the difficulty lies in evidence reasoning rather than task-format understanding.

Fine-Tuning Performance Fine-tuning produces the strongest results. Qwen2.5-14B achieves AUC-ROC 70.4%, followed by Qwen3-14B (69.5%), OpenBioLLM-8B (67.5%), and Llama-3.1-8B (66.9%), all exceeding Bashir et al. (2021) (63.6%). The difference between Qwen2.5-14B and the random forest baseline is significant (McNemar's $p = 0.0015$). At threshold 0.37, Qwen2.5-14B achieves 65.7% balanced accuracy with 63.6% Changed recall and 67.9% specificity.

At matched scale, OpenBioLLM-8B (67.5%) exceeds Llama-3.1-8B (66.9%), indicating a modest gain from biomedical pretraining. The gap to Qwen2.5-14B (70.4%) is therefore primarily attributable to scale rather than domain mismatch. Context length is a secondary factor: OpenBioLLM's 8,192-token pretraining range is exceeded by most instances with 21+ publications (Appendix B), whereas the 14B Qwen models natively support our 16,384-token input. Disentangling these effects would require a matched biomedical 14B model with equivalent context, which is not currently available.

7 Influence of Publication Count

At threshold 0.37, the fine-tuned model produces 164 true positives, 285 true negatives, 94 false negatives, and 135 false positives. False negatives involved significantly fewer new publications than true positives (median 3 vs. 16; Mann–Whitney $U = 1{,}794$, $p < 0.001$), while false positives involved more than true negatives (median 12 vs. 2; $U = 35{,}630$, $p < 0.001$): the model over-predicts

Table 2: Test-set performance ($n = 678$; 420 Unchanged, 258 Changed), reported as % with 95% bootstrap CIs (2,000 resamples). †Threshold selected on training data to maximise balanced accuracy; others use the default boundary. **Bold**: best standard fine-tuned model; underline: best zero-shot.

		Overall						Unchanged			Changed		
Method	**Model**	**Acc**	**BalAcc**	**AUC**	**P_M**	**R_M**	**$F1_M$**	**P_U**	**R_U**	**$F1_U$**	**P_C**	**R_C**	**$F1_C$**
Simple Baselines													
Majority class	—	61.95	50.00	50.00	30.98	50.00	38.24	61.95	100.0	76.50	0.00	0.00	0.00
Random (50/50)	—	50.00	50.00	50.00	50.00	50.00	43.22	61.95	50.00	55.37	38.05	50.00	43.22
Prior Work (Bashir et al., 2021; reimplemented on our dataset)													
LR	4 features	61.21 [57.5,64.9]	59.50 [55.6,63.5]	60.43 [55.8,64.9]	59.28 [55.5,63.2]	59.50 [55.6,63.5]	59.35 [55.5,63.2]	69.48 [65.0,73.7]	66.67 [62.3,71.3]	68.04 [64.2,71.7]	49.09 [43.4,55.3]	52.33 [46.3,58.4]	50.66 [45.4,56.1]
DT	4 features	55.75 [51.9,59.6]	59.28 [55.5,62.9]	59.76 [55.6,63.9]	59.33 [55.5,62.9]	59.28 [55.5,62.9]	55.75 [51.9,59.5]	73.62 [67.8,78.9]	44.52 [39.6,49.1]	55.49 [50.7,59.8]	45.05 [40.8,49.9]	74.03 [68.4,79.3]	56.01 [51.8,60.4]
RF	4 features	59.00 [55.3,62.8]	59.65 [55.9,63.6]	63.62 [59.4,68.0]	59.10 [55.5,62.8]	59.65 [55.9,63.6]	58.45 [54.7,62.3]	71.13 [66.2,75.9]	56.90 [52.1,61.7]	63.23 [59.1,67.1]	47.08 [42.2,52.5]	62.40 [56.7,68.2]	53.67 [49.1,58.5]
This Work — Zero-Shot General LLMs													
Standard (P-01)	Qwen2.5-14B	52.51 [48.7,56.2]	55.69 [51.8,59.4]	60.53 [56.2,64.9]	55.69 [51.8,59.4]	55.69 [51.8,59.4]	52.51 [48.6,56.2]	68.99 [63.0,74.8]	42.38 [37.7,47.1]	52.51 [47.6,57.0]	42.38 [37.5,47.2]	68.99 [63.1,74.6]	52.51 [47.7,56.9]
Debiased (P-02)	Qwen2.5-14B	47.05 [43.2,50.7]	54.05 [50.9,57.1]	61.10 [56.7,65.4]	55.62 [51.2,59.8]	54.05 [50.9,57.1]	45.59 [41.7,49.2]	70.75 [62.9,78.2]	24.76 [20.6,29.0]	36.68 [31.5,41.7]	40.49 [36.4,44.6]	83.33 [78.8,87.8]	54.50 [50.2,58.4]
Tiered (P-03)	Qwen2.5-14B	54.42 [50.6,58.3]	54.54 [50.8,58.4]	58.56 [54.3,62.8]	54.28 [50.7,58.0]	54.54 [50.8,58.4]	53.70 [49.9,57.5]	66.18 [61.2,71.1]	54.05 [49.4,58.8]	59.50 [55.3,63.5]	42.39 [36.9,47.8]	55.04 [48.7,61.3]	47.89 [42.8,52.6]
Standard (P-01)	GPT-4.1-mini	55.60 [51.9,59.3]	57.96 [54.2,61.6]	62.86 [58.6,67.1]	57.70 [54.1,61.2]	57.96 [54.2,61.6]	55.53 [51.8,59.2]	70.88 [65.2,75.9]	48.10 [43.4,53.1]	57.30 [52.8,61.7]	44.53 [39.8,49.4]	67.83 [61.9,73.5]	53.76 [49.1,58.1]
Debiased (P-02)	GPT-4.1-mini	52.36 [48.7,56.2]	59.53 [56.7,62.4]	63.35 [59.1,67.6]	62.98 [59.3,66.8]	59.53 [56.7,62.4]	51.14 [47.5,55.0]	82.12 [75.7,88.1]	29.52 [25.5,33.9]	43.43 [38.5,48.4]	43.83 [39.7,48.0]	89.53 [85.8,93.3]	58.85 [54.6,62.9]
Tiered (P-03)	GPT-4.1-mini	60.62 [56.6,64.5]	56.85 [53.2,60.6]	61.83 [57.6,66.2]	57.35 [53.4,61.4]	56.85 [53.2,60.6]	56.91 [53.0,60.8]	66.74 [62.2,71.1]	72.62 [68.0,76.9]	69.56 [65.8,72.9]	47.96 [41.4,54.7]	41.09 [35.3,47.2]	44.26 [38.8,49.8]
This Work — Few-Shot													
3-shot random	Qwen2.5-14B	51.03 [47.3,54.7]	54.27 [50.4,57.9]	58.96 [54.9,63.4]	54.30 [50.4,57.9]	54.27 [50.4,57.9]	51.03 [47.3,54.7]	67.32 [61.4,73.0]	40.71 [36.1,45.4]	50.74 [46.2,55.1]	41.27 [36.6,46.0]	67.83 [61.8,73.5]	51.32 [46.6,55.9]
This Work — Fine-Tuned													
FT	Qwen2.5-14B+QLoRA	**68.44** [64.7,72.1]	63.83 [60.3,67.4]	**70.36** [66.3,74.3]	**66.38** [62.3,70.6]	63.83 [60.3,67.4]	64.17 [60.3,68.0]	70.93 [66.7,75.0]	83.10 [79.4,86.7]	**76.54** [73.3,79.5]	**61.83** [54.4,69.0]	44.57 [38.4,50.6]	51.80 [46.0,57.3]
FT†	Qwen2.5-14B+QLoRA	66.22 [62.5,69.6]	**65.71** [61.9,69.3]	—	65.02 [61.3,68.4]	**65.71** [61.9,69.3]	**65.11** [61.2,68.5]	**75.20** [70.6,79.3]	67.86 [63.5,72.3]	71.34 [67.7,74.7]	54.85 [49.2,60.4]	**63.57** [57.4,69.2]	**58.89** [53.7,63.5]
FT	Qwen3-14B+QLoRA	67.55 [63.9,70.9]	62.67 [59.0,66.1]	69.49 [65.3,73.4]	65.32 [61.0,69.4]	62.67 [59.0,66.1]	62.90 [58.9,66.6]	70.08 [66.1,74.0]	83.10 [79.5,86.7]	76.03 [72.9,79.0]	60.56 [52.8,67.6]	42.25 [36.3,48.4]	49.77 [43.8,55.4]
FT†	Qwen3-14B+QLoRA	65.34 [61.7,68.9]	64.17 [60.3,67.8]	—	63.74 [60.0,67.2]	64.17 [60.3,67.8]	63.86 [60.0,67.4]	73.42 [69.2,77.7]	69.05 [64.9,73.6]	71.17 [67.6,74.6]	54.06 [48.2,59.9]	59.30 [53.4,65.2]	56.56 [51.3,61.1]
FT	Llama-3.1-8B+QLoRA	62.83 [59.3,66.4]	52.13 [50.4,54.0]	66.86 [62.5,70.8]	61.19 [52.2,69.9]	52.13 [50.4,54.0]	44.73 [41.7,48.1]	63.00 [59.3,66.6]	96.90 [95.2,98.5]	76.36 [73.5,79.1]	59.38 [42.3,76.0]	7.36 [4.3,10.8]	13.10 [7.9,18.5]
FT†	Llama-3.1-8B+QLoRA	64.75 [60.9,68.4]	57.79 [54.6,61.1]	—	61.92 [57.2,66.5]	57.79 [54.6,61.1]	56.79 [52.7,60.8]	66.48 [62.5,70.4]	86.90 [83.6,90.0]	75.34 [72.2,78.3]	57.36 [48.6,65.4]	28.68 [23.2,34.3]	38.24 [31.8,44.2]
FT	OpenBioLLM-8B+QLoRA	66.22 [62.5,69.9]	60.33 [56.8,63.9]	67.50 [63.2,71.5]	63.81 [59.2,68.3]	60.33 [56.8,63.9]	60.13 [55.9,64.2]	68.26 [64.3,72.2]	85.00 [81.5,88.4]	75.72 [72.5,78.8]	59.35 [51.2,67.1]	35.66 [29.5,41.8]	44.55 [38.0,50.6]
FT†	OpenBioLLM-8B+QLoRA	65.78 [62.1,69.3]	63.93 [60.1,67.5]	—	63.81 [59.9,67.4]	63.93 [60.1,67.5]	63.87 [60.0,67.4]	72.71 [68.3,76.9]	71.67 [67.4,75.9]	72.18 [68.5,75.6]	54.92 [48.8,60.8]	56.20 [50.2,62.1]	55.56 [50.2,60.4]

change when many publications are present and under-predicts when few are present.

Overall calibration is good (Expected Calibration Error = 0.028, 10 bins), so the probabilities reflect genuine confidence rather than calibration artefacts. However, predictions above 0.80 are poorly calibrated (19 instances; 58% accuracy vs. 91% mean confidence), consistent with overconfidence in the high-publication regime.

Table 3 stratifies Changed-class recall and Unchanged-class specificity by the number of new publications. When only 1–2 new publications are available, the model detects just 18.4% of conclusion changes. Recall rises monotonically with publication count, reaching 95.6% for instances with 21 or more publications, while specificity shows the inverse pattern, dropping from 95.9% to 2.9%. This indicates that the model uses publication count as a proxy for conclusion change, rather than reasoning about evidence content. The predicted probability $P(\text{CHANGED})$ correlates with publication count at Spearman $\rho = 0.80$, far exceeding the true association between publication count and the label ($\rho = 0.30$). Thus, the model has amplified a moderate real association into a dominant decision rule.

Pubs.	*n*	Changed	Recall	Spec.
1–2	219	49	18.4	95.9
3–5	149	50	30.0	87.9
6–10	115	47	72.3	47.1
11–20	92	44	93.2	4.2
21+	103	68	95.6	2.9
All	678	258	63.6	67.9

Table 3: Changed-class recall and Unchanged-class specificity of the fine-tuned model (threshold 0.37), stratified by the number of new publications.

Although publication count is never an explicit feature, the concatenated input exposes it through two channels: positional encoding of the prediction token (which scales with input length) and textual cues such as repeated abstract structures across publications. To isolate the positional channel, we repeated inference with all inputs left-padded to 16,384 tokens, eliminating variation in the prediction token's position. The correlation was virtually unchanged ($\rho = 0.78$ padded vs. $\rho = 0.80$ unpadded; Pearson $r = 0.99$ between probability vectors), confirming that the model infers volume from textual cues rather than from sequence position alone. These cues may be reinforced by attention patterns that concentrate on tokens near sequence boundaries (Liu et al., 2024), though we do not test this directly.

7.1 Publication Count Ablation at Inference

To test whether performance depends on evidence volume or content, we provide only the first k publication abstracts at inference; all other settings are identical. If the model reasons about content, performance should be largely preserved at small k. Table 4 presents the results. At k=1, AUC-ROC drops from 70.4% to 53.4%, barely above chance, and Changed-class recall falls to 5.8%. Performance improves monotonically with k, and the shortcut correlation ρ rises from -0.26 at k=1 to 0.80 at k=All. These results establish a relationship between evidence volume and model performance: this shortcut is the primary driver of the fine-tuned model's discrimination.

k	AUC-ROC	BalAcc	Chg Rec	Spec	ρ
1	53.4 [48.9, 57.8]	51.7	5.8	97.6	−0.26
3	59.3 [54.9, 63.6]	53.7	12.4	95.0	0.05
5	62.3 [58.0, 66.4]	54.5	20.9	88.1	0.36
10	68.9 [64.7, 72.9]	64.7	59.3	70.0	0.73
All	70.4 [66.3, 74.3]	65.7	63.6	67.9	0.80

Table 4: Publication count ablation at inference. The fine-tuned model (Qwen2.5-14B, threshold 0.37) is evaluated with only the first k publication abstracts provided at inference; all other settings are identical.

7.2 Publication Subsampling

To explore the effect of removing publication count information, we trained Qwen2.5-14B with counterfactual publication subsampling at two levels: k=1 and k=3 publications per training instance. For each training instance, k publications were randomly sampled (or all, when fewer than k existed), decoupling publication count from the label during training; at inference all publications were used, leaving the test distribution unchanged. Other hyperparameters were unchanged.

Table 5 presents the results alongside the standard model (k=All). A clear trade-off emerges across three dimensions: overall discrimination (AUC-ROC), shortcut strength (ρ), and low-publication recall.

At k=1, where each training instance contains a single publication abstract, the shortcut is nearly eliminated: ρ drops from 0.80 to 0.15, below even the true label–publication count association

Method	AUC	ρ	1–2 Rec	BalAcc
FT k=1 subsample	55.4	0.15	57.1	52.0
FT k=3 subsample	63.5	0.36	46.9	59.4
FT k=All (standard)	70.4	0.80	18.4	65.7

Table 5: Shortcut–discrimination trade-off. AUC-ROC (%), Spearman ρ between predicted probability and publication count, Changed-class recall (%) at 1–2 publications, and balanced accuracy (%) at threshold 0.37 (selected on the training set).

($\rho = 0.30$). Changed-class recall at 1–2 publications rises from 18.4% to 57.1%, a threefold improvement in the clinically critical low-publication regime. However, overall AUC-ROC falls to 55.4%, barely above chance, indicating the model has lost the volume signal without learning to compensate through content-based reasoning.

At k=3, the trade-off is more balanced: $\rho = 0.36$ (close to the true association), 1–2 publication recall improves to 46.9%, and AUC-ROC remains at 63.5%, comparable to the Bashir et al. random forest baseline (63.6%).

8 Conclusion

We introduced conclusion change prediction as a multi-document evidence reasoning task on 3,326 Cochrane review-update pairs. Fine-tuning achieves the strongest overall discrimination (AUC-ROC 70.4%) but analysis indicates that it relies heavily on an evidence volume shortcut which future work should aim to overcome.

Limitations

The dataset is restricted to Cochrane intervention reviews with binary labels. Models are provided with abstracts, keeping the system broadly deployable given uneven full-text access. Publications were identified retrospectively and retrieval noise in deployment would likely degrade accuracy. Fine-tuning used a single A100 80 GB GPU with 4-bit quantisation, and 154 long-input instances (Appendix B) were right-truncated. Single training run per configuration.

Ethics Statement

All information required to reproduce our work is contained in this paper: dataset construction and label derivation procedures are described in Section 3; model training configurations and hyperparameters in Section 4.3; the full text of all prompts in Appendix A; and input length and truncation details in Appendix B. Researchers with Cochrane Library access can reconstruct the dataset and reproduce the experiments from these specifications.

Acknowledgements

We gratefully acknowledge the Cochrane Library for providing access to the systematic review data used in this work.

References

Amal Alharbi and Mark Stevenson. 2019. A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1257–1260.

Rabia Bashir, Adam G Dunn, and Didi Surian. 2021. A rule-based approach for automatically extracting data from systematic reviews and their updates to model the risk of conclusion change. *Research Synthesis Methods*, 12(2):216–225.

Rabia Bashir, Didi Surian, and Adam G Dunn. 2019. The risk of conclusion change in systematic review updates can be estimated by learning from a database of published examples. *Journal of Clinical Epidemiology*, 110:42–49.

Aaron M Cohen, Kyle Ambert, and Marian McDonagh. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making*, 12(1):33.

Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Miranda Cumpston and Ella Flemyng. 2023. Chapter IV: Updating a review. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions, version 6.4*. Cochrane.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. *Advances in Neural Information Processing Systems*, 36.

Julian H Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, and 1 others. 2017. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91:23–30.

Paul Garner, Sally Hopewell, Jackie Chandler, Harriet MacLehose, Elie A Akl, Joseph Beyene, Stephanie Chang, Rachel Churchill, Karin Dearness, Gordon Guyatt, and 1 others. 2016. When and how to update systematic reviews: consensus and checklist. *BMJ*, 354.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Diana Herrera-Perez, Alyson Haslam, Tyler Crain, Jennifer Gill, Catherine Livingston, Victoria Kaestner, Michael Hayes, Dan Morgan, Adam S Cifu, and Vinay Prasad. 2019. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *eLife*, 8:e45183.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, 4(1):78.

Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B. Hugging Face model repository.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How quickly do systematic reviews go out of date? a survival analysis. *Annals of Internal Medicine*, 147(4):224–233.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, and 1 others. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1):55.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Prompt Templates

This appendix provides the full text of all prompts used in our experiments. All prompts instruct the model to output a single token: `N0` (unchanged) or `N1` (changed). The user message is identical across all variants and consists of the review abstract followed by the concatenated publication abstracts.

A.1 Zero-Shot Prompts (Section 4.1)

P-01 (Standard). This prompt provides a straightforward task description with no additional guidance.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

P-02 (Debiased). This prompt adds base-rate information and an explicit instruction to counteract majority-class bias.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

IMPORTANT: In this dataset, approximately 38% of systematic reviews have changed conclusions upon updating. Do NOT default to predicting unchanged. Carefully assess whether the new evidence warrants a change.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

P-03 (Tiered). This prompt asks the model to reason in two explicit steps before making a prediction.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

Reason in two steps:
Step 1: Assess whether any new publication reports findings that conflict with, contradict, or substantially extend the review's existing conclusions.
Step 2: Based on your assessment, decide whether the overall conclusions would need to change.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

A.2 Fine-Tuning Prompt (Section 4.3)

The fine-tuning prompt is shown in Figure 2 in the main text. It is reproduced verbatim below for completeness.

```
You are an expert biomedical evidence analyst. Your job is to decide if the conclusions of a systematic review (SR) should be updated in light of new trial evidence. The publication abstracts provided below have already been screened by expert reviewers and confirmed as relevant to this review. All publications meet the inclusion criteria. Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

The user message follows the template:

```
SYSTEMATIC REVIEW AND NEW TRIAL ABSTRACTS:
{review_abstract}
{publication_abstracts}
Decision:
```

During training, the assistant message contains the gold label token (`N0` or `N1`). During inference, the assistant message is omitted and the model generates its prediction.

A.3 Few-Shot Prompt (Section 4.2)

The few-shot prompt uses the same system message as P-01 (Standard). Three training examples

are prepended to the user message as demonstrations. Each demonstration consists of a review–publication pair followed by the gold label.

```
Here are three examples of this task:

– Example 1 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_1_text}
Decision: N1

– Example 2 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_2_text}
Decision: N0

– Example 3 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_3_text}
Decision: N0

Now predict for this instance:

SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{test_instance_text}
Decision:
```

For each test instance, the three exemplars are randomly sampled from the training set.

The user message is identical to all other methods. Probability estimates were obtained by extracting logit values for the decision token from the classification call and normalising via softmax.

B Input Length and Truncation

Table 6 reports the distribution of input token counts and truncation at the 16,384-token limit, stratified by the number of new publications per instance. Truncation is concentrated at the high-publication tail.

Pubs.	n	Median tok.	Truncated	% Trunc.
1–2	1,055	1,254	0	0.0
3–5	757	2,269	0	0.0
6–10	625	3,903	0	0.0
11–20	482	6,699	1	0.2
21–50	329	12,819	75	22.8
51+	78	31,320	78	100.0
All	3,326	2,738	154	4.6

Table 6: Input length and truncation at the 16,384-token limit, stratified by the number of new publications. "% Trunc." is the proportion of instances in the row whose tokenised length exceeds 16,384 tokens, at which point later publications are right-truncated. All instances with 51 or more new publications exceed the limit; none with 10 or fewer do. Overall, 95.4% of instances fit within the limit without truncation.

VaxScope: Document-Level Structured Evidence Extraction from Immunization Systematic Reviews

Bahar İlgen[1], Ebenezer Awotoro[1, 2], Georges Hattab[1,2]
[1] Centre for Artificial Intelligence in Public Health Research (ZKI-PH),
Robert Koch Institute, Berlin, Germany
[2] Department of Mathematics and Computer Science,
Freie Universität Berlin, Berlin, Germany
ilgenb@rki.de

Abstract

Systematic reviews are fundamental to evidence-based medicine, but the clinical evidence they contain is primarily expressed in unstructured text, making large-scale extraction and reuse difficult. Existing biomedical NLP methods have achieved strong performance on span-level extraction from clinical trials and abstracts; however, these approaches are insufficient for systematic reviews, where evidence is often distributed across multiple studies, sentences, and sections and must be aggregated into normalized document-level attributes. We introduce **VaxScope**, a benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews. VaxScope is constructed through an expert-guided semi-automatic annotation pipeline that combines automatic candidate generation with domain expert validation to ensure consistency and annotation quality. We formalize the task as document-level structured extraction, where target labels are defined at the review level and require aggregating evidence beyond isolated textual spans. We further establish baselines for document-level structured extraction using abstract-level input representations and evaluate how access to evidence-grounded contextual input improves performance over abstract-only settings. Baseline experiments show that PubMedBERT achieves the best overall performance (Avg F1: 0.850), with evidence-grounded input improving performance particularly for fields requiring distributed contextual reasoning.

1 Introduction

Systematic reviews are central to evidence-based medicine, but the evidence they contain is primarily expressed in unstructured natural language, making large-scale extraction, comparison, and reuse difficult for downstream applications such as clinical decision support and evidence synthesis (Pilic et al., 2023). Existing biomedical NLP methods have achieved strong performance on patient-level clinical text, such as electronic health records and clinical notes (Lee et al., 2019; Alsentzer et al., 2019). They have also shown strong results in span-level extraction from biomedical abstracts and trial reports (Fraile Navarro et al., 2023). However, these approaches are often insufficient for systematic reviews, where evidence is synthesized across multiple studies and relevant information is distributed across sentences and document sections rather than localized in isolated spans.

This limitation is particularly important in the immunization domain, where vaccine recommendations and public health policies depend heavily on systematic reviews that summarize evidence across populations, pathogens, interventions, and clinical outcomes (Pilic et al., 2023; İlgen et al., 2024). Supporting decisions in this setting requires structured access not only to explicit entities such as diseases and populations, but also to higher-level attributes such as outcomes, study characteristics, evidence summaries, and review design.

A central challenge is that many of these attributes cannot be reliably extracted from individual text spans alone (Zhang et al., 2024; Schmidt et al., 2025). Instead, they require aggregating evidence across multiple mentions and resolving information at the document level (Zhang et al., 2020; Zheng et al., 2024). For example, the study design or the number of included studies may be mentioned in different sections, while clinically relevant outcomes are often described implicitly across multiple paragraphs. As a result, standard span-based extraction approaches alone are insufficient to capture the full structure of evidence represented in systematic reviews.

To address this limitation, we introduce **VaxScope**, a benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews. Each review is annotated with a set of normalized attributes, in-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 853–863
July 3-4, 2026 ©2026 Association for Computational Linguistics

cluding population, disease, outcome, study design, number of included studies, and review type. In addition, each structured annotation is linked to a small set of supporting evidence snippets, enabling traceability, human verification, and evidence-grounded extraction and validation. This evidence-grounded design supports interpretable downstream applications and human-in-the-loop decision support (Riskin et al., 2025; DeYoung et al., 2020).

VaxScope is constructed using an expert-guided semi-automatic annotation pipeline that combines automatic candidate generation with domain expert validation to ensure consistency and annotation quality. We establish benchmark baselines for document-level structured extraction using abstract-level input, and demonstrate that multi-task transformer classifiers effectively capture heterogeneous document-level attributes across categorical, multi-label, and numeric fields. These structured attributes can additionally support interactive and visualization-based systems for exploring immunization evidence (İlgen and Hattab, 2025). Taken together, these design choices position VaxScope as both a benchmark dataset and a practical resource for downstream evidence exploration. Our contributions are threefold:

- We introduce **VaxScope**, a new benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews.
- We design an evidence-grounded annotation schema that combines normalized document-level attributes with linked supporting evidence snippets, enabling traceable, interpretable, and verifiable extraction.
- We establish benchmark baselines for document-level structured extraction using abstract-level input, and evaluate how access to evidence-grounded contextual input improves performance relative to abstract-only settings.

2 Related Work

Biomedical natural language processing has made substantial progress in extracting structured information from clinical and scientific text (Lee et al., 2019). Early approaches relied on rule-based methods using predefined lexical and syntactic patterns, which proved effective for high-precision extraction in constrained settings but often struggled with ambiguous and heterogeneous biomedical language (Wang et al., 2017). More recently, transformer-based models pretrained on biomedical corpora, such as BioBERT (Lee et al., 2019) and BIO_CLINICALBERT (Alsentzer et al., 2019), have substantially improved performance across named entity recognition, relation extraction, and document classification tasks. However, much of this progress has focused on sentence-level or span-level extraction from primary clinical documents, leaving document-level structured evidence extraction from systematic reviews comparatively underexplored.

Multi-label classification of biomedical documents has been studied extensively in the context of literature indexing and topic annotation. Du et al. (2019) proposed ML-Net, a deep learning architecture for multi-label biomedical text classification that jointly predicts labels and label counts. More recently, the BioCreative VII LitCovid track (Chen et al., 2022; Lin et al., 2022) introduced a shared task for multi-label topic annotation of COVID-19 literature, demonstrating the utility of transformer-based classifiers for document-level label assignment in rapidly evolving biomedical domains. While these tasks are closely related to the multi-label components of our framework, they primarily focus on topical document classification. In contrast, VaxScope requires heterogeneous document-level extraction that jointly combines categorical labels, multi-label aggregation, and structured numeric and temporal fields.

A prominent line of work in biomedical NLP focuses on extracting structured evidence from randomized controlled trial (RCT) reports using the PICO framework—Population, Intervention, Comparator, and Outcome. The EBM-NLP corpus (Nye et al., 2018) provided a foundational benchmark for span-level PICO extraction from RCT abstracts and has been widely used to evaluate sequence labeling and NER-based approaches. Subsequent work has explored multi-task learning, section-aware models, and prompt-based methods to improve extraction precision (Hu et al., 2023). More recently, large language models have been applied to PICO extraction in zero-shot and few-shot settings (Ghosh et al., 2024; Chen et al., 2025), achieving competitive performance without requiring labeled training data. However, these approaches remain predominantly span-level and are designed for primary study reports (Lehman et al., 2019). In contrast,

VaxScope targets structured extraction from *systematic reviews*, where evidence is synthesized across multiple primary studies and key attributes often require document-level aggregation rather than isolated span detection.

Automating systematic reviews has attracted growing interest as the volume of biomedical literature continues to expand (Marshall and Wallace, 2019; Tsafnat et al., 2014). Prior work has addressed several stages of the systematic review pipeline, including relevant paper retrieval (van de Schoot et al., 2021), risk of bias assessment (Gates et al., 2018), and evidence synthesis and summarization (Wang et al., 2022). Recent studies have explored the use of LLMs for semi-automated data extraction from systematic reviews, reporting strong performance for extracting predefined entities and study characteristics across multiple domains (Schmidt et al., 2024). At the same time, recent living systematic reviews of systematic review automation show that most existing approaches remain centered on extracting PICO-style evidence from primary studies and randomized controlled trials, while structured extraction of review-level metadata remains comparatively underexplored (Schmidt et al., 2025). In contrast, VaxScope focuses on the review itself as the primary unit of analysis: given a systematic review document, the goal is to extract normalized attributes that characterize the review as a whole, including its scope, methodology, and evidence base. While VaxScope captures methodological attributes such as study design, it does not currently include explicit evidence-quality assessments such as GRADE scores or risk-of-bias ratings. These dimensions fall outside the scope of the current abstract-based benchmarking and typically require a broader full-text methodological appraisal beyond the extraction targets considered here. Incorporating such fields remains an important direction for future dataset extensions.

Standard information extraction systems operate at the sentence or span level, identifying entities and relations within individual sentences. However, in many real-world settings, relevant information is distributed across multiple sentences and sections of a document and cannot be reliably captured through isolated span extraction (Zhang et al., 2020). Document-level relation extraction has emerged as an active research area, with methods that model cross-sentence dependencies through graph neural networks and transformer architectures (Zhou et al., 2021; Yao et al., 2019). In the biomedical domain, document-level extraction has been studied for chemical–disease relation extraction (Li et al., 2016) and other entity-level tasks. VaxScope extends this perspective to structured attribute extraction from systematic reviews, where multiple heterogeneous fields—spanning categorical labels, multi-label sets, and numeric or temporal values—must be jointly extracted at the document level. This heterogeneous formulation distinguishes our task from prior document-level extraction benchmarks, which typically focus on a single relation type or entity category.

3 Methods

3.1 Dataset Construction Pipeline

The VaxScope corpus was constructed through a semi-automatic data creation pipeline designed to balance scalability and annotation quality. We first collected immunization-related systematic reviews and meta-analyses from an expert-curated collection of immunization evidence. For each document, candidate annotations were identified from the full text, as relevant information is often distributed across multiple sections, including methods, results, and discussion. We then applied automatic candidate extraction to identify preliminary values for structured fields such as the number of included studies, date of the last literature search, review type, and disease mentions. For the initial seed set, automatically extracted candidates were subsequently reviewed and corrected through expert validation. This semi-automatic pipeline enabled efficient corpus construction while preserving annotation consistency and clinical relevance, as illustrated in Figure 1.

An existing domain-expert annotated benchmark set of 100 immunization-related systematic reviews served as the foundation for schema design, protocol validation, and final evaluation. We used this benchmark to derive label categories, normalization rules, omission rules, and evidence selection criteria, which were then formalized into a constrained prompt schema for LLM-assisted candidate generation. The model was therefore not used to define labels or annotation logic, but to apply an expert-derived annotation protocol at scale. Candidate generation was performed using `gemini-3-flash-preview` (Google DeepMind), prompted with a constrained JSON schema that defined normalized value spaces, controlled vocab-

ularies, and critical extraction rules for each field (e.g., `date_of_last_lit` was set to `null` if not explicitly stated in the text, to avoid unsupported date assignment). The full prompt schema is provided in Appendix A. Generated candidates were subsequently reviewed and corrected where necessary, while the final evaluation was conducted exclusively on the independently expert-annotated gold set.

Following this expert annotation protocol, the corpus was expanded using an LLM-assisted semi-automatic pipeline to 2,000 annotated documents. During data inspection, we identified 79 duplicate review instances between the expert benchmark and the larger LLM-assisted corpus. To avoid evaluation leakage, these overlapping documents were removed before model development. The remaining corpus consisted of 1,921 documents, which were split into 1,536 training and 385 development instances.

Annotation quality was assessed through two complementary steps: (1) automatic quality evaluation by comparing LLM-generated candidate annotations against the expert gold set, reporting field-level F1 and exact-match accuracy; and (2) inter-annotator agreement analysis on a randomly sampled subset of 50 documents from the expanded corpus. These documents were independently annotated by two of the authors following the expert-derived annotation protocol and without access to the original LLM-generated labels. Agreement was measured using Cohen's κ for categorical fields, Jaccard similarity for multi-label fields, and exact-match accuracy for numeric and temporal fields, as reported in Table 1.

3.2 Task Definition

We formulate the problem as document-level structured evidence extraction from systematic reviews in the immunization domain. Given a review document (typically represented by its abstract), the goal is to map the document to a set of normalized attributes that summarize its key evidence characteristics. Although candidate annotations are identified from the full text during corpus construction, the abstract is used as the primary input for the extraction models, as it provides a consistent and practically deployable representation across studies. This setting differs fundamentally from standard span-based information extraction tasks. In systematic reviews, relevant information is heterogeneous and distributed across multiple sentences and

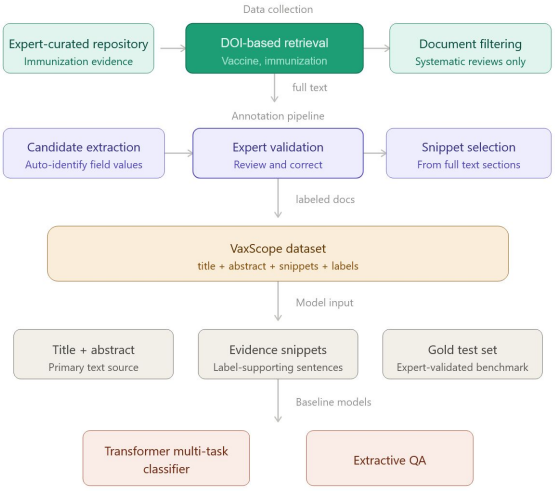


Figure 1: Semi-automatic corpus construction pipeline for VaxScope. Documents are collected from an expert-curated immunization evidence source and retrieved via DOI lookup. Candidate annotations are generated automatically and validated by domain experts. Evidence snippets from the full text provide support for document-level labels.

sections, and cannot be reliably captured through isolated span extraction. For example, population characteristics may appear in the inclusion criteria, outcomes in the results, and study characteristics in the methods. Extracting structured information, therefore, requires consolidating evidence that is often distributed across multiple mentions within a document.

We define the task as predicting a structured representation comprising predefined fields: population, outcome, disease, number of included studies, study design, and review type. These fields correspond to different types of prediction problems: some require normalized categorical classification (e.g., disease, review type), others require aggregating evidence across multiple mentions (e.g., population, study design), and others involve extracting normalized numeric or temporal values (e.g., number of studies, date of last literature search).

To support interpretability and annotation consistency, each document is additionally associated with a small set of evidence snippets (typically a few sentences) that jointly support the extracted attributes, providing traceable justification for document-level annotations and facilitating human verification. In the current benchmark setting, these evidence snippets are included in the annotation schema and used by the baselines as contextual input. However, the current baselines do not ex-

Field	Metric	Score
Disease	Cohen's κ	0.84
Review type	Jaccard	0.89
Num. studies	EM / Near-miss	0.90 / 0.92
Num. participants	EM	0.82
Date of last search	EM / Year-level	0.82 / 0.90
Topic	Jaccard	0.76
Outcome	Jaccard	0.69
Population	Jaccard	0.71

Table 1: Inter-annotator agreement on 50 randomly sampled documents. Cohen's κ is used for disease, Jaccard similarity for multi-label fields, and exact match (EM) with near-miss tolerance for numeric and temporal fields. Near-miss denotes ± 1 for num. studies and year-level match for date of last search. For the outcome field, agreement was computed only over instances where both annotators assigned a non-null value, as null entries reflected topic-conditional field activation rather than outcome-category decisions.

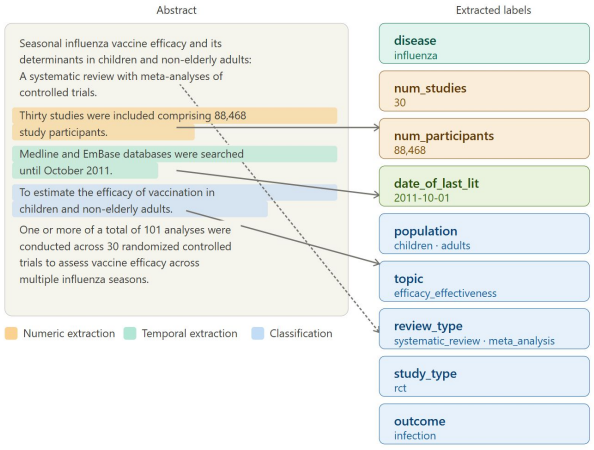


Figure 2: Example annotation from VaxScope. Highlighted spans indicate evidence snippets supporting extracted labels. Amber marks numeric fields, green temporal fields, and blue categorical and multi-label fields. Evidence snippets may be drawn from the abstract or full-text sections depending on where supporting evidence appears in the source document.

plicitly predict snippet selection within their output space.

3.3 Annotation Schema

We define a document-level annotation schema that captures key structured attributes of systematic reviews in the immunization domain. Each document is mapped to a set of predefined fields representing clinically and methodologically relevant aspects of the evidence, such as population, outcome, disease, study design, and review type. Figure 2 shows an annotated example from the VaxScope corpus, illustrating how evidence snippets support the extracted structured labels.

All fields are annotated at the document level, supported by evidence snippets drawn from across the full review. To ensure consistency, each field is associated with a defined value space, including normalized categories (e.g., population groups, study designs) or structured values (e.g., numeric counts and dates). Importantly, the schema is heterogeneous: different fields correspond to different types of prediction problems, including categorical classification, multi-label aggregation over distributed mentions, and numeric or temporal slot filling. This design reflects the diverse structure of information in systematic reviews and aligns with the task formulation described in the previous section. Table 2 summarizes the annotation schema, including field definitions, prediction types, and value spaces. In addition to structured fields, each document is linked to a small set of evidence snippets that provide traceable support for the annotations. The schema is designed to balance expressiveness and the feasibility of annotations. While some fields are normalized into predefined categories, others allow flexible values to capture domain-specific variations. This design supports both structured modeling and practical annotation at scale.

3.4 Baseline Models

We implement a hybrid extraction architecture that combines slot-filling extractors for numeric and temporal fields with a transformer-based multi-task classifier for categorical and multi-label document-level attributes. Figure 3 illustrates the overall extraction architecture.

Multi-task transformer classifier. For categorical and multi-label document-level attributes—including disease, review type, topic, study design, population, and outcome—we evaluate three transformer-based baselines with task-specific prediction heads: BIO_CLINICALBERT and PUBMEDBERT as biomedical encoder baselines, and CLINICAL-LONGFORMER as a long-context baseline. In the primary experimental setting, the model takes the document title and abstract as input and produces a shared document-level representation for downstream field-specific prediction heads. To assess the added value of evidence-aware contextualization, we additionally report an oracle upper-bound setting in which expert-selected evidence snippets are concatenated with the title

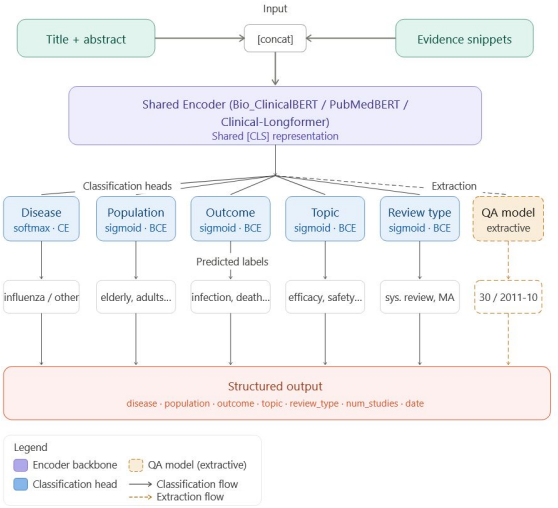


Figure 3: Architecture of the VaxScope extraction system. The shared encoder produces a shared document-level representation that is passed to task-specific classification heads for categorical and multi-label fields. Numeric and temporal fields are extracted separately using an extractive QA model (dashed).

and abstract as additional contextual input. Single-label fields (e.g., disease) are modeled using softmax classification with cross-entropy loss, while multi-label attributes (e.g., population, outcome) are modeled using sigmoid output layers with binary cross-entropy loss.

Numeric and temporal extraction. For fields requiring numeric or temporal values—specifically the number of included studies, the total number of participants, and the date of the last literature search—we adopt an extractive QA approach using `deepset/roberta-base-squad2`, a RoBERTa-based (Liu et al., 2019) question-answering model fine-tuned on SQuAD 2.0 (Rajpurkar et al., 2018). For each target field, we define a natural language question (e.g., *"How many studies were included in this systematic review?"*) and extract the answer span from the document context. Extracted spans are post-processed into typed values (integers or normalized dates in `YYYY-MM-DD` format). Answers below a confidence threshold are mapped to `null`. To validate this design choice, we also compare against a deterministic **regex baseline** that uses regular expressions and domain-specific lexical patterns. This approach can achieve high precision when target values are expressed in predictable surface forms, and serves as a lower-bound reference for the slot-filling fields.

3.5 Evaluation Metrics

Categorical and multi-label fields are evaluated using field-level precision, recall, and F1 score. For numeric slot-filling fields (e.g., number of included studies and total number of participants), we report exact-match accuracy and near-miss accuracy with ± 1 tolerance. This is particularly important in systematic reviews, where multiple related numeric quantities may co-occur within the same document (e.g., numbers screened, excluded, and included), making target disambiguation non-trivial. For temporal fields, we report exact-match accuracy on normalized date strings (`YYYY-MM-DD`) as well as partial-match accuracy at the year level.

4 Experiments

4.1 Experimental Setup

Dataset. VaxScope comprises 1,921 documents drawn from immunization-related systematic reviews, split into 1,536 training and 385 development instances. A held-out set of 100 expert-validated documents is reserved exclusively for final test evaluation.

Models. We evaluate three transformer-based encoders as a shared backbone for the multi-task classifier: BIO_CLINICALBERT (Alsentzer et al., 2019), a biomedical encoder pretrained on clinical notes; PUBMEDBERT (Gu et al., 2021), pretrained exclusively on PubMed abstracts; and CLINICAL-LONGFORMER (Li et al., 2022), a long-context encoder supporting up to 4,096 tokens.

Input settings. Although the labels are defined at the review level and require document-level interpretation, the current baselines use abstract-level input as a standardized and practically accessible benchmark setting. We evaluate two input configurations: a primary setting using only the document title and abstract, and an oracle setting that additionally incorporates expert-selected evidence snippets drawn from the full text. The primary setting reflects a realistic deployment scenario in which only abstract-level information is available. The oracle setting assumes ideal evidence localization and serves as an upper bound on extraction performance. This comparison isolates the contribution of explicit evidence grounding to document-level structured extraction.

Hyperparameters. All models are fine-tuned for 12 epochs using AdamW (lr=2×10^{-5}, weight de-

Field name	Description	Prediction type	Value / label space
Number of studies included	Number of primary studies included in the review	Numeric slot filling	Single integer (e.g., 30)
Date of last literature search	Date of the most recent literature search	Temporal slot filling	Normalized date (stored as YYYY-MM-DD) or null
Review type	Type of review	Multi-label classification	systematic_review, meta_analysis, rapid_review, living_review, umbrella_review, other
Study design	Study designs of included primary studies	Multi-label aggregation	rct, cohort, case_control, cross_sectional, case_series, qualitative, mixed_method, other
Disease	Target disease or pathogen of the review	Single-label classification	Controlled vocabulary (e.g., influenza, covid_19, . . .)
Population	Target population groups	Multi-label aggregation	children, adolescents, adults, elderly, pregnant, healthcare_workers, immunocompromised, all age groups, other
Outcome	Main clinical outcomes for effectiveness analyses	Multi-label classification	infection, hospitalization, death, other
Topic (evidence type)	Primary evidence focus of the review	Multi-label classification	efficacy_effectiveness, immunogenicity, safety, . . .

Table 2: Document-level annotation schema in the VaxScope corpus. All fields are extracted at the document level and supported by selected evidence snippets. Value spaces show representative categories; the full label space is defined in the annotation schema provided with the dataset.

cay=0.01). BIO_CLINICALBERT and PUBMEDBERT use batch size 4 and maximum sequence length 512. CLINICAL-LONGFORMER uses batch size 2 and maximum sequence length 2,048. A sigmoid threshold of 0.5 is applied for multi-label classification.

4.2 Results

Table 3 presents development set results across all model and input setting combinations. We report Macro-F1 for disease classification and Micro-F1 for all multi-label fields.

4.3 Analysis

PUBMEDBERT achieves the best overall performance across both input settings, reaching Avg F1 of 0.797 in the abstract-only setting, 0.810 with evidence snippets on the development set, and 0.850 on the held-out gold test set. This suggests that domain-specific pretraining on PubMed abstracts provides a stronger inductive bias than long-context modeling for this task, likely because systematic review abstracts are structurally dense and domain-specific. CLINICAL-LONGFORMER performs competitively in the abstract-only setting but benefits more from evidence snippets, particularly for *outcome* and *study_type*, suggesting that longer context is most useful when distributed evidence is explicitly provided. Evidence snippets consistently improve performance across models, with

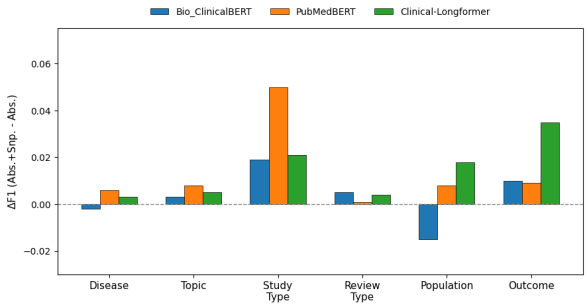


Figure 4: Performance gain from adding expert-selected evidence snippets (Abs.+Snp.) over abstract-only input (Abs.) across prediction fields. Values show ΔF1 on the development set. Gains are largest for study type, population, and outcome, while disease and review type show minimal improvement.

the largest gains observed for *study_type*, *population*, and *outcome*, while *disease* and *review_type* show minimal improvement due to their reliance on standardized labels that are already identifiable from abstracts. The *population* field remains the most challenging across all settings, consistent with inter-annotator agreement results (Jaccard = 0.71), likely reflecting label overlap and implicit population mentions in immunization literature. The *outcome* field also shows relatively lower agreement (Jaccard = 0.69), consistent with its topic-conditional nature.

Model	Setting	Disease	Topic	Study Type	Review Type	Population	Outcome	Avg
Majority baseline	–	0.105	0.324	0.375	0.781	0.284	0.356	0.371
Bio_ClinicalBERT	Abs.	0.945	0.777	0.714	0.948	0.580	0.628	0.765±0.007
Bio_ClinicalBERT	Abs.+Snp.	0.943	0.780	0.733	0.953	0.565	0.638	0.769±0.008
PubMedBERT	Abs.	0.952	0.799	0.752	0.957	0.623	0.698	0.797±0.011
PubMedBERT	Abs.+Snp.	0.958	0.807	0.802	0.958	0.631	0.707	0.810±0.009
Clinical-Longformer	Abs.	0.951	0.785	0.717	0.950	0.591	0.674	0.778±0.009
Clinical-Longformer	Abs.+Snp.	0.954	0.790	0.738	0.954	0.609	0.709	0.792±0.003

Table 3: Development set F1 scores. Macro-F1 is reported for disease, and Micro-F1 for all other fields. Abs. denotes title + abstract only; Abs.+Snp. denotes title + abstract with expert-selected evidence snippets.

Model	Abs.	Abs.+Snp.
Bio_ClinicalBERT	0.806±0.008	0.817±0.009
PubMedBERT	0.838±0.002	0.850±0.006
Clinical-Longformer	0.758±0.001	0.773±0.046

Table 4: Final evaluation results on the held-out expert gold test set. "Abs." uses title + abstract only, while "Abs.+Snp." additionally includes expert-selected evidence snippets. Values report Avg F1, computed as the unweighted mean across task-specific scores (Macro-F1 for disease; Micro-F1 for all remaining fields).

Field	Method	EM	Near-Miss
Num. studies	Regex	0.290	0.290
	QA	0.290	0.290
	QA+Snp.	0.470	0.480
Num. participants	Regex	0.360	0.360
	QA	0.440	0.440
	QA+Snp.	0.600	0.600
Date of last search	Regex	0.020	0.192
	QA	0.172	0.364
	QA+Snp.	0.293	0.576

Table 5: Slot-filling results on the held-out gold test set. EM: exact match; Near-miss: ±1 for numeric fields, year-level for temporal fields. Only documents with non-null gold annotations are evaluated.

5 Discussion

VaxScope addresses a gap in biomedical NLP by providing a benchmark for document-level structured evidence extraction from systematic reviews. Our results show that transformer-based multi-task classifiers can effectively extract heterogeneous structured attributes from immunization-related systematic reviews, achieving strong performance on well-defined fields such as disease and review type. A key finding is that domain-specific pretraining (PUBMEDBERT) is more beneficial than long-context modeling (CLINICAL-LONGFORMER), suggesting that the structural density of systematic review abstracts limits the practical benefit of extended context windows, while familiarity with PubMed-style language provides a stronger inductive bias for this task. Evidence snippets consistently improve performance for fields requiring distributed contextual reasoning, particularly study type, outcome, and population, while highly standardized fields such as disease and review type remain robust under abstract-only input. Lower performance on numeric and temporal slot-filling fields (Table 5) should be interpreted less as model failure and more as a consequence of the intentionally abstract-based benchmark setting, since values such as the number of included studies and the date of the last literature search are often absent from abstracts and only reliably reported in full-text methods sections. In this sense, snippet gains function as indicators of full-text dependency rather than simple performance improvements: large gains for population, outcome, and study type reflect distributed evidence requirements, whereas low-gain fields such as disease and review type are largely recoverable from abstract-only input. Inter-annotator agreement further confirms that well-defined categorical fields such as disease and review type can be annotated consistently, whereas population and outcome remain more challenging due to label ambiguity and implicit mentions. This is also reflected in model performance, where population and outcome remain the most challenging fields across all settings.

6 Conclusion

We introduced VaxScope, a dataset and benchmark for document-level structured evidence extraction from immunization-related systematic reviews. Our baseline experiments show that multi-task transformer classifiers achieve strong performance across heterogeneous document-level fields, with PUBMEDBERT achieving the best overall result (Avg F1: 0.850 on the held-out gold test set). Overall, VaxScope highlights the importance of document-level modeling for structured evidence extraction and provides a foundation for automated evidence synthesis in public health. The VaxScope dataset and annotation schema are available at https://github.com/baharilgen/VaxScope.

Limitations

The current baselines leverage long-context self-attention to capture cross-sentence dependencies implicitly; however, they do not explicitly model mention-level conflict resolution or evidence graphs. More structured aggregation mechanisms, such as hierarchical sentence encoders or graph-based evidence modeling, remain important directions for future work. Furthermore, VaxScope is currently limited to the immunization domain, and generalization to other systematic review domains remains to be explored. In addition, although evidence snippets are included to support interpretability and downstream traceability analyses, the current baselines do not explicitly predict or evaluate snippet selection. Snippet-level grounding and joint label–evidence evaluation also remain important directions for future work.

Acknowledgments

The authors thank the Immunization Unit (FG33) at the Robert Koch Institute for providing the expert-annotated reference set and access to the SYSVAC immunization review registry that supported corpus construction and evaluation.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj, Jingcheng Du, Li Fang, Kai Wang, Shuo Xu, Yuefu Zhang, Parsa Bagherzadeh, Sabine Bergler, Aakash Bhatnagar, Nidhir Bhavsar, Yung-Chun Chang, Sheng-Jie Lin, Wentai Tang, Hongtong Zhang, Ilija Tavchioski, Senja Pollak, and 20 others. 2022. Multi-label classification for biomedical literature: an overview of the biocreative vii litcovid track for covid-19 literature topic annotations. *Database*, 2022:baac069.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280. Open access.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

Allison Gates, Ben Vandermeer, and Lisa Hartling. 2018. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the robotreviewer machine learning tool. *Journal of Clinical Epidemiology*, 96:54–62.

Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. Alpapico: Extraction of pico frames from clinical trial documents using llms. *Methods*, 226:78–88.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Yan Hu, Vipina K. Keloth, Kalpana Raja, Yong Chen, and Hua Xu. 2023. Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics*, 39(9):btad542.

Bahar İlgen and Georges Hattab. 2025. Visualizing evidence through natural language interaction for public health interventions. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 431–438, New York, NY, USA. Association for Computing Machinery.

Bahar İlgen, Antonia Pilic, Thomas Harder, and Georges Hattab. 2024. Pre-training to identify immunization-related entities from systematic reviews. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '23, page 234–239, New York, NY, USA. Association for Computing Machinery.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.

2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *Preprint*, arXiv:2201.11838.

Sheng-Jie Lin, Wen-Chao Yeh, Yu-Wen Chiu, Yung-Chun Chang, Min-Huei Hsu, Yi-Shin Chen, and Wen-Lian Hsu. 2022. A bert-based ensemble learning approach for the biocreative vii challenges: full-text chemical identification and multi-label classification in pubmed articles. *Database*, 2022:baac056.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Iain J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8:163.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Antonia Pilic, Sarah Reda, Catherine L Jo, Helen Burchett, Magdalena Bastías, Pauline Campbell, Deepa Gamage, Louise Henaff, Benjamin Kagina, Wiebe Külper-Schiek, Carole Lunny, Melanie Marti, Rudzani Muloiwa, Dawid Pieper, James Thomas, Matthew C Tunis, Zane Younger, Ole Wichmann, and Thomas Harder. 2023. Use of existing systematic reviews for the development of evidence-based vaccination recommendations: Guidance from the sysvac expert panel. *Vaccine*, 41(12):1968–1978.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Daniel Jay Riskin, Keri L. Monda, Joshua J. Gagne, Robert Reynolds, A. Reshad Garan, Nancy Dreyer, Paul Muntner, and Brian D. Bradbury. 2025. Implementing accuracy, completeness, and traceability for data reliability. *JAMA Network Open*, 8(3):e250128.

Lena Schmidt, Kaitlyn Hair, Sergio Graziozi, Fiona Campbell, Claudia Kapp, Alireza Khanteymoori, Dawn Craig, Mark Engelbert, and James Thomas. 2024. Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. In *Proceedings of the 3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (ALTARS 2024)*, Glasgow, UK. CEUR Workshop Proceedings.

Lena Schmidt, Ailbhe N. Finnerty Mutlu, Rebecca Elmore, Babatunde K. Olorisade, James Thomas, and Julian P. T. Higgins. 2025. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. *F1000Research*, 10:401. Originally published 2021-05-19; Version 3 published 2025-04-08.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic Reviews*, 3(1):74. Open access commentary.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3:125–133.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2017. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale

document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Gongbo Zhang, Yiliang Zhou, Yan Hu, Hua Xu, Chunhua Weng, and Yifan Peng. 2024. A span-based model for extracting overlapping pico entities from randomized controlled trial publications. *Journal of the American Medical Informatics Association*, 31(5):1163–1171.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hanwen Zheng, Sijia Wang, and Lifu Huang. 2024. A comprehensive survey on document-level information extraction. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 58–72, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.

A Annotation Prompt Schema

The following pseudo-JSON schema was used to prompt `gemini-3-flash-preview` (Google DeepMind) for candidate annotation generation, based on an expert-derived annotation protocol.

```
/* Pseudo-JSON annotation schema */
{
  "doc_id": "doc_XXXX",
  "doi": "...",
  "title": "...",
  "abstract": "...",
  "disease": "...",
  "population": [...],
  "topic": [...],
  "outcome": [...],
  "study_type": [...],
  "review_type": [...],
  "num_studies": null,
  "num_participants": null,
  "date_of_last_lit": null,
  "evidence_snippets": [...]
}
```

A.1 Critical Extraction Constraints

The following constraints were enforced to ensure annotation consistency and reduce schema-inconsistent outputs:

- `date_of_last_lit`:
  set to `null` if not explicitly stated in the text. Publication date must not be used as a substitute.
- `population`:
  if all ages are included or unspecified, use `["all_age_groups"]` alone; no other values may co-occur.
- `disease`:
  use `vaccine_preventable_diseases` only when no single primary disease can be identified.
- `outcome`:
  set to `null` when `topic` does not include `efficacy_effectiveness`; otherwise, one or more of: `infection`, `hospitalization`, `death`, `other`.
- `evidence_snippets`:
  a small set of supporting sentences copied verbatim from the source document; paraphrasing or summarization is not permitted.

A.2 Example Annotation

The following example is drawn from the expert-annotated gold set.

```
{
  "doc_id": "doc_0023",
  "doi": "10.1136/bmjopen-2017-019206",
  "title": "Parents' uptake of human
      papillomavirus vaccines for their
      children: a systematic review and
      meta-analysis of observational
      studies",
  "disease": "human_papillomavirus",
  "population": ["parents_caregivers"],
  "topic": ["coverage", "acceptance"],
  "outcome": null,
  "study_type": ["cross_sectional", "
      cohort", "case_control"],
  "review_type": ["systematic_review", "
      meta_analysis"],
  "num_studies": 79,
  "num_participants": 840838,
  "date_of_last_lit": "2017-11-30",
  "evidence_snippets": [
    "Seventy-nine studies on 840 838
        parents across 15 countries were
         included.",
    "The pooled proportion of parents'
        uptake of HPV vaccines for their
         children was 41.5% (range:
        0.7%-92.8%), twofold higher for
        girls (46.5%) than for boys
        (20.3%).",
    "...",
  ]
}
```

Medical Context Variation: A source of impairment for Event classification

Aman Sinha[1,2,3] **Marianne Clausel**[1] **Mathieu Constant**[2] **Xavier Coubez**[3]

[1]IECL, University of Lorraine, France
[2]ATILF, CNRS, Université de Lorraine, Nancy, France
[3]Institut Strauss, Strasbourg, France
Correspondence: aman.sinha@univ-lorraine.fr

Abstract

The variation in writing style encapsulates nuanced characteristics, which are often exploited for author or demographic identification. In the medical domain, language models are frequently deployed to capture relevant information from unstructured or complex data, such as clinical notes that often include patients' medical histories. Such data is largely free-form and unstructured, obtained through diverse clinician–patient interactions. In this work, we present a case study investigating whether variations in clinicians' writing styles can lead to differences in medical context understanding capabilities for pre-trained language models (PLMs) on five downstream tasks related to medical event classification. Our findings indicate that variation in writing style, characterized by linguistic features, can indeed lead to suboptimal performance in deployed NLP models. Furthermore, we explore linguistically guided counterfactual reasoning to mitigate the impact of writing-style variation, suggesting that LLM-based stylistic normalization is effective for this purpose.

1 Introduction

Clinical narrative plays important role in determining the history of the patient. It helps clinicians in answering prognostic inquiries and is essential for identifying the confounding between disease progression and interventions (Pham et al., 2017). The documentation writing practices are designed to be chronological and divided into identifiable sections, such as patient's prior medical context, physical examination, possible diagnosis and actions (Bansler et al., 2016). Therefore, the narrative found in clinical notes can depend on the writing style of the clinical staff. Furthermore, the variation in writing can be a source of communicating incomplete, ambiguous or at times wrong information.

For eg., In the sentence – *The infusion rate was then increased and blood was taken 4,8,12 and 20 minutes after the new target concentration had been achieved.* The writer were interested in how quickly the concentration reached its new steady-state after the rate of infusion had been increased; they took blood 4, 8, 12 and 20 minutes after taking the action that sought to produce a new target concentration, that is, after altering the rate of infusion. The investigators could not know the concentration of the drug in the blood until after analysis some days later. A recommended alternative (Goodman and Edwards, 2014) for the text is as follows – *The infusion rate was then increased and blood was taken 4, 8, 12 and 20 minutes later, or (although it is implicit and not strictly necessary) ... after the new target concentration had been set.* This sentence clarifies the order of the actions and the events happening.

English for Medical Purposes (EMP) is a key focus in medical education, with research emphasizing the importance of collocation acquisition by EFL learners (Džuganová, 2019). Medical writing often suffers from issues such as overly complex sentences, excessive passive voice, weak verb usage, and medical jargon, all of which hinder clarity and comprehension (Goodman and Edwards, 2014). Given that free-text clinical notes make up about 80% of clinical data, models trained on such text are likely influenced by stylistic and demographic variation. While large-scale training may offer robustness to writing style differences, this assumption has not been systematically evaluated for language models.

In this work, we present a case study focusing on linguistic variation, in particular writing style variation, that is found in the clinical notes for the task of extracting contextual information related to medicinal changes. The primary objective is to investigate *whether the characteristics of writing style in clinical notes has subsequent impact on language models* for event classification tasks. Our key contributions in this paper are as follows:

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 864–879
July 3-4, 2026 ©2026 Association for Computational Linguistics

(a) We provide a thorough linguistic feature analysis to find empirical evidence for stylistic variation in clinical notes has impact on capability of pre-trained language models.
(b) We show that linguistically-guided counterfactual reasoning can be used to mitigate the impact of writing style variation on LMs.
(c) Additionally, we study the generated counterfactuals for medical relevance for data augmentation.

2 Background

Style Variation Style variation is a form of linguistic variation has been studied for describing and explaining variation in linguistic form across society (Nguyen et al., 2021). Stylistic Vagueness or variation is very well acknowledged in medical writing (Mitchell, 2007) and is advised to avoid as much as possible in medical communication. Although, the existence of any linguistic variation can be attributed to demographics of international doctors across the world, majority of whom have English as their second-language (Verma et al., 2016). Linguistic variation between user and the content of clinical notes already has studied for medical dialog systems (Llanos et al., 2016). However, the impact of the writing style variation has not been studied thoroughly for language models for medical context understanding.

Medical Contextual Understanding Clinical notes are narrative and longitudinal, capturing a patient's medical history, assessments, and care plans across time. These texts vary widely in structure and writing style depending on the provider, institution, and note type (Pivovarov and Elhadad, 2015). Events such as medication changes are often embedded in nuanced context — temporal cues, attribution, and intent — that must be accurately captured for tasks like medication reconciliation (Cadwallader et al., 2013). While clinical event extraction has received significant attention, the effect of writing-style variation on the medical contextual understanding of language models remains underexplored (Xu et al., 2022).

Counterfactual Explanations We consider counterfactual explanation (Wachter et al., 2017; Sinha et al., 2023) as a mean to verify and potentially mitigate the impact the different linguistic feature that contribute to writing style variation in clinical notes on medical language models' capabilities for contextual medical event extraction task. Various works that exist inline with this approach such as POLYJUICE (Wu et al., 2021), MICE (Ross et al., 2021) which provide a pipeline for counterfactual generation and evaluation. However, there is a lack of exploration into mitigating medical writing style variation, which our work addresses.

3 Methodology

Dataset CMED (Mahajan et al., 2022) dataset is a collection of clinical notes that are associated to the patients who were admitted for the Type II Diabetes (T2D). The dataset consists of total 500 clinical notes from two different hospitals (See Appendix A for more details). Each clinical note was annotated by a team of three annotators led by a physician for medication events extraction and context annotations.

Contextual Event Extraction The contextual medical event extraction was introduced in n2c2 challenge[1]. It involved two-fold tasks, firstly, it requires to identify medicine for which any *medication change*[2] is being discussed in the clinical note and secondly, different contextual information associated the medication change has to be extracted. For any medication change that is identified as *disposition*, the model is expected to extract five different contextual information regarding the `Action`, `Negation`, `Temporality`, `Certainty` and `Actor` that are associated to for each identified medication change (Mahajan et al., 2022). See Appendix A for more details.

Linguistic Characterization of Clinical Narrative We utilize BEA linguistic feature extractor tool (Lee and Lee, 2023) in order to extract linguistic features from any segmented medical sentence of the clinical notes. The tool facilitates to extract 220 linguistic feature per sentence (See Table 19,20). These features cover various linguistic characteristics on sentence-level such as surface, syntax, lexico-semantics, and discourse. We utilized SciSpacy[3] instead of general domain modules for processing the segmented sentences.

Models We consider four pre-trained language models as follows: BERT (Devlin et al., 2018),

[1] https://n2c2.dbmi.hms.harvard.edu/

[2] Medication change, is defined as "any discussion about a medicine change" and is termed to as *disposition*, for a given patient mentioned in the clinical note.

[3] en_core_sci_sm

Linguistic Feature (*lf*)	**Medical Sentence**	**Modifed Sentence**	**BEA**(s)	**BEA**(s_l)
#unique_nouns	–check chem 7 given recent restart BP meds	–check chemistry 7 given recent restart blood pressure medications	3	5
simple_verbs_variation	5 x2 and lopressor 12.	5 multiplied by 2 and lopressor 12.	0	1
#unique_adverbs	Pain was relieved by Morphine (4, 4, 2 mg).	Pain was completely relieved by Morphine (4, 4, 2 mg).	0	1
root_verbs_variation	8. Neurontin 200 mg b.i.d. (she was told she could discontinue	Neurontin 200 mg b.i.d. (she was informed she could cease taking it).	1.414	1.732
#named_entities_person	We will have her start the Norvasc and Cozaar, have her return in two weeks for followup of blood pressure.	We will have Jane start the Norvasc and Cozaar, have Jane return in two weeks for followup of blood pressure.	0	2
simple_type_token_ratio	She was treated with IV lopressor 2.	She received IV lopressor twice.	1.125	1.2
avg_syllables_per_word	A1c off it 3/94 8.	A1c off it three over ninety-four eight.	0.667	1.25

Table 1: Samples showing linguistic features modification. BEA(..) denotes linguistic feature counter.

lf	D	$D(l)$	**Action**	**Actor**	**Certainty**	**Negation**	**Temporality**
#unique_nouns	2.94	3.55	neg.	pos.	pos.	pos.*	pos.
simple_verbs_variation	0.89	0.95	pos.*	neg.	neg.	pos.*	neg.
#unique_adverbs	0.60	1.04	pos.*	pos.	pos.	pos.*	pos.
root_verbs_variation	1.23	1.34	neg.*	pos.	pos.	neg.*	pos.
#named_entities_person	0.27	0.65	/	pos.	/	/	/
simple_type_token_ratio	1.10	1.13	neg.*	pos.*	pos.	neg.*	pos.
avg_syllables_per_word	1.41	1.67	pos.*	pos.	neg.*	pos.*	neg.

Table 2: Mean shift for selected linguistic features and their correlation impact (pos./neg.) on context categories. * denotes non-significant, / denotes no correlation. D: original dataset and $D(l)$: modified dataset.

You are an assistant supporting a linguistic study. Your task is to rewrite the given sentence by modifying the following linguistic feature: [feature]. Return only the modified sentence, without additional text or explanation. Sentence: "sentence"

Table 3: Prompt used for generating sentence explanations.

ModernBERT (Warner et al., 2024), ClinicalBERT (Huang et al., 2019) and PubmedBERT (Gu et al., 2021). For each of these BERT-like models, we train classification model using the segmented sentences containing the target medicine as an input against each of the five context category labels for 20 epochs with early stopping. Details for reproducibility in Appendix C.

4 Experimental Setup

Each clinical note is segmented at the sentence level, and then sentences containing a medication change annotated with a *disposition* label are retained. We train the different LMs to obtain predictions on the development and test sets. Additionally, we extract 220 linguistic features using BEA linguistic feature extraction library for selected disposition based sentences. To test our hypothesis, we first define sample-level *correctness* by comparing ensemble predictions of all the language models with gold labels. Then, with the obtained ensemble predictions, we examine correlations between linguistic features and correctness via the OLS method (Kuchibhotla et al., 2019) across the five context categories.

Based on obtained correlation analysis, we select a subset of features and list them with their expected impact in Table 2. Next, we use GPT-3.5-Turbo to modify validation and test segmented sentences to alter linguistic features for counterfactual evaluation using the prompt (See Table 3). We provide examples of modified sentence with respect to a set of linguistic features in Table 1). Finally, the modified sentences are used for inference with the trained LMs to verify the if the counterfactual are able to change the predictions or not as expected.

5 Results & Discussion

Table 4 shows the percentage of linguistic features (lf) that were significantly correlated with prediction mistakes in dev and test set across the five context tasks (See details in Table 12 - 18). We consider seven most common linguistic features for which we obtain *llm* modification across the five context tasks. Table 1 presents examples of linguistic modifications generated via GPT-3.5-Turbo for each of the selected linguistic feature. Additionally, we measure and provide the overall feature shift shown in Table 2 as modified dataset compared to the original average lf distribution. Table 5 presents the detailed results on original and modified dataset for test set. We also report broader linguistic categorization for each selected features based on (Lee and Lee, 2023) documentation in Table 11.

Finding I: Writing style variation impacts models' ability to understand clinical narratives.

CONTEXT	Action	Actor	Certainty	Negation	Temporality
DEV	3.64	42.27	27.73	12.27	67.73
TEST	11.82	37.73	21.36	7.73	9.55

Table 4: % of linguistic features that are significantly (ρ<0.05) correlated with prediction mistakes.

Among the five context tasks, *Actor*, *Certainty*, and *Temporality* are most affected by variation in linguistic features (see Table 4). In the test set, *Negation* appears to be least influenced by linguistic features, whereas in the development set, *Action* showed the least impact, indicating inconsistency across datasets. As shown in Table 11, lexico-semantic and syntactic feature families account for the majority of significant correlations with model prediction correctness, suggesting that models may be relying on surface stylistic cues rather than meaningful medical context distinctions. The exception is observed for *Actor* and *Negation*, where discourse-related linguistic features play a dominant role. Overall, these findings reinforce that writing style variation substantially shapes how models interpret and process clinical narratives.

Models	Action	Actor	Certainty	Negation	Temporality
BERT	51.84	44.70	52.80	49.54	48.30
ModernBERT	43.55	35.83	42.54	63.90	53.58
BioClinicalBERT	49.13	50.38	48.77	49.54	50.12
PubmedBERT	71.48	49.95	53.65	49.54	39.37
ENSEMBLE	52.50	50.20	50.20	49.50	48.50
#unique_nouns	-0.7	+1.0	-1.9	±0.0	+1.5
simple_verbs_variation	+2.1	-6.3	+5.2	±0.0	+0.3
#unique_adverbs	-1.0	+0.9	-1.9	±0.0	-1.7
root_verbs_variation	-0.7	-3.0	+0.6	±0.0	-1.3
#named_entities_person	-0.3	-2.2	±0.0	±0.0	+0.1
simple_type_token_ratio	-0.2	-6.8	+3.8	±0.0	+1.5
avg_syllables_per_word	-2.7	-4.1	-1.2	±0.0	+2.6

Table 5: Ensemble context classification results (Macro F1) under counterfactual feature shift on the test set. Green indicates expected outcomes, red indicates unexpected outcomes, and gray denotes excluded cases.

Finding II: Linguistically-guided counterfactuals show potential to mitigate impact of writing style variation. Compared to ensemble baseline performance across all the five tasks, counterfactual modified dataset leads to both gains and losses depending on the linguistic feature are modified (See Table 5). Complete results are in the Appendix B. Concretely, for overall ∼60% cases the modified dataset via counterfactuals leads to the expected behavior. *Negation* remains relatively invariant across all feature shifts, however for *Actor*, *Certainty* and *Temporality*, we obtain mixed behavior. *Actor*, *Certainty* are the most affected showing significant performance change under *simple_verbs_variation* and *simple_type_token_ratio* respectively. Conversely, *Action* and *Negation* seems to experience least changes on when subjected to linguistic feature shift. These results suggest that further confirms that linguistic variations can both degrade and enhance contextual understanding suggesting a more non trivial phenomena.

	Modification Type	Example
LOW	Underspecification	This is rather a simple infection, and may be treated with Ancef. This is rather a simple cellulitis, and may be treated with Ancef.
MODERATE	Overspecification	Reinitiated Lasix promptly. Reinitiated Lasix.
	Entity	She was started on Xenical by two doctors. She was started on Xenical by Internal Medicine.
	Coreference	insulin dosing per Dr. Nettles. Dr. Nettles has been eating and drinking insulin dosing per Dr. Nettles. She has been eating and drinking
HIGH	Content Violation	No John with Mary given possible Sarah bleed. No heparin with HD given possible GI bleed.
	Numeric	2 liters of IV fluid, came up to 90s/50s. 5 liters of IV fluid, came up to 90s/50s.

Table 6: Different LLM generated counterfactuals categorized based on medical risk.

lf	LOW	MODERATE	HIGH
#unique_nouns	79.6	10.2	10.2
simple_verbs_variation	65.3	14.3	20.4
#unique_adverbs	89.8	8.2	2.0
root_verbs_variation	98.0	0.0	2.0
#named_entities_person	65.3	14.3	20.4
simple_type_token_ratio	95.9	2.0	2.0
avg_syllables_per_word	93.9	6.1	0.0

Table 7: Categorization of Counterfactual Explanation based on clinical relevance risk.

Finding III: Linguistic-guided counterfactuals are double-edged sword. The use of LLM to generate counterfactuals by default is prone to produce hallucinations. Therefore, we analyse the modified sentences (see Table 1) for hallucination to understand the clinical relevance distortion [4]. We consider 50 segmented sentences that are modified with respect to seven linguistic features via GPT-3.5-Turbo. We manually annotate each of the modified sentence (50x7=350) for the level of potential medical relevance distortion risk into three categories: low, moderate, and high (see Table 6). For LOW risk, we consider modifications that involve under-specification of any linguistic unit i.e., plausible but unsupported modifiers. For MODER-

[4]Medical relevance distortion denotes to what extent the counterfactual deviates from original sentence on clinical ground.

ATE risk, we consider modification of clinically relevant content while remaining linguistically plausible. For HIGH risk, we consider violations of transformation constraints.

Among the seven features we utilized for generating counterfactuals, based on 50 instances, we observed that *named_entities_nouns*, *unique_nouns* leads to high risk counterfactuals whereas *root_verbs_variation* leads to low risk counterfactuals. We present in distribution in Table 7.While HIGH risk modifications, due to distortion of medical relevance, can mislead the models to deviate from intended downstream tasks. LOW and MODERATE level modifications can be useful to augment the model to overcome class imbalance or lack of clinical annotated data.

6 Conclusion

In this study, we investigated the impact of writing style variation on medical context classification task performance of pre-trained language models. Our findings indicate that stylistic variation in clinical narratives can lead to difference in medical context interpretation of language models. Our study highlights the impact of considering writing style heterogeneity, suggesting that incorporating stylistic normalization via counterfactual explanation generation could enhance model robustness. While the counterfactuals explanation can often require clinical validation, otherwise, it can positively contributes to address the lack of clinical annotated data.

Limitations

While the study highlights the impact of writing style variation on language models in medical domain, we acknowledge the following limitations to the best of our knowledge.

- Firstly, the study focuses primarily on linguistic features and does not account for other potential sources of variation, such as differences in clinical knowledge or documentation practices.
- Secondly, our case study examines the effect on selected downstream tasks from correlation angle, and therefore findings may not generalize to all possible confounders.
- And finally, our study focuses on pre-trained language models considering practical deployment settings and encourage future studies to generalize the finding of LLMs.

Acknowledgments

We would like to extend our thanks to the anonymous reviewers. Additionally, we would also like to express thanks to Cristina Garcia Holgado and Timothée Mickus for the valuable discussion on the initial version of this work.

References

Jørgen Bansler, Erling Havn, Kjeld Schmidt, Troels Mønsted, Helen Petersen, and Jesper Svendsen. 2016. Cooperative epistemic work in medical practice: An analysis of physicians' clinical notes. *Computer Supported Cooperative Work*, 25.

Justin Cadwallader, Kenneth Spry, Justin Morea, AL Russ, Jon Duke, and Michael Weiner. 2013. Design of a medication reconciliation application. *Applied clinical informatics*, 4(01):110–125.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Božena Džuganová. 2019. Medical language-a unique linguistic phenomenon: Engleski. *Jahr–European Journal of Bioethics*, 10(1):129–145.

Neville W Goodman and Martin B Edwards. 2014. *Medical writing: a prescription for clarity*. Cambridge University Press.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Arun K Kuchibhotla, Lawrence D Brown, Andreas Buja, and Junhui Cai. 2019. All of linear regression. *arXiv preprint arXiv:1910.06386*.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Leonardo Campillos Llanos, Dhouha Bouamor, Pierre Zweigenbaum, and Sophie Rosset. 2016. Managing linguistic and terminological variation in a medical

dialogue system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3167–3173.

Diwakar Mahajan, Jennifer J. Liang, and Ching-Huei Tsou. 2022. Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *AMIA Annual Symposium Proceedings*, 2021:833–842.

SLM Mitchell. 2007. Medical writing: A prescription for clarity.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 603–612.

Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Aman Sinha, Cristina Garcia Holgado, Marianne Clausel, Mathieu Constant, and Xavier Coubez. 2023. Can llms be used to understand clinical notes better? *5èmes journées du Groupement de Recherche CNRS «Linguistique Informatique, Formelle et de Terrain» LIFT 2023*, page 94.

Anju Verma, Ann Griffin, Jane Dacre, and Andrew Elder. 2016. Exploring cultural and linguistic influences on clinical communication skills: a qualitative study of international medical graduates. *BMC Medical Education*, 16(1):162.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Zhuo Xu, Yue Wang, Lu Bai, and Lixin Cui. 2022. Writing style aware document-level event extraction. *arXiv preprint arXiv:2201.03188*.

A More Information for CMED Dataset

The CMED dataset contains clinical notes that came from the Partners HealthCare Electronic Medical Records (EMR). EMR at Partners HealthCare comprises a platform shared by two large academic tertiary hospitals – Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), USA. The five context categories description are below:

Action	What is the change discussed?
Actor	Who initiated the change?
Certainty	How likely is this change to have occurred / will occur?
Negation	Is the change being discussed negated?
Temporality	When is this change intended to occur?

Table 8: Description of Medical contextual event classification tasks.

The distribution of classes across the different context tasks for train set is shown in Figure 1.

B Detailed Test Results

Models	Action	Actor	Certainty	Negation	Temporality
BERT	51.84	44.70	52.80	49.54	48.30
ModernBERT	43.55	35.83	42.54	63.90	53.58
BioClinicalBERT	49.13	50.38	48.77	49.54	50.12
PubmedBERT	71.48	49.95	53.65	49.54	39.37
ENSEMBLE	52.50	50.20	50.20	49.50	48.50
#unique_nouns	51.80	51.20	48.30	49.50	50.00
simple_verbs_variation	54.60	43.90	55.40	49.50	48.80
#unique_adverbs	51.50	51.10	49.50	49.50	46.80
root_verbs_variation	51.80	47.20	50.80	49.50	47.20
#named_entities_person	52.20	48.00	50.20	49.50	48.60
simple_type_token_ratio	52.30	434.00	54.00	49.50	50.00
avg_syllables_per_word	49.80	46.10	49.00	49.50	51.10

Table 9: Ensemble context classification results (Macro F1) under counterfactual feature shift on the test set (Complete Results).

Table 9 presents the experiment results in detail. The top section contains the five context categories across all four PLMs, along with the ENSEMBLE results. The bottom section comprises ensemble results corresponding to linguistic features listed in each row when the models are trained with a

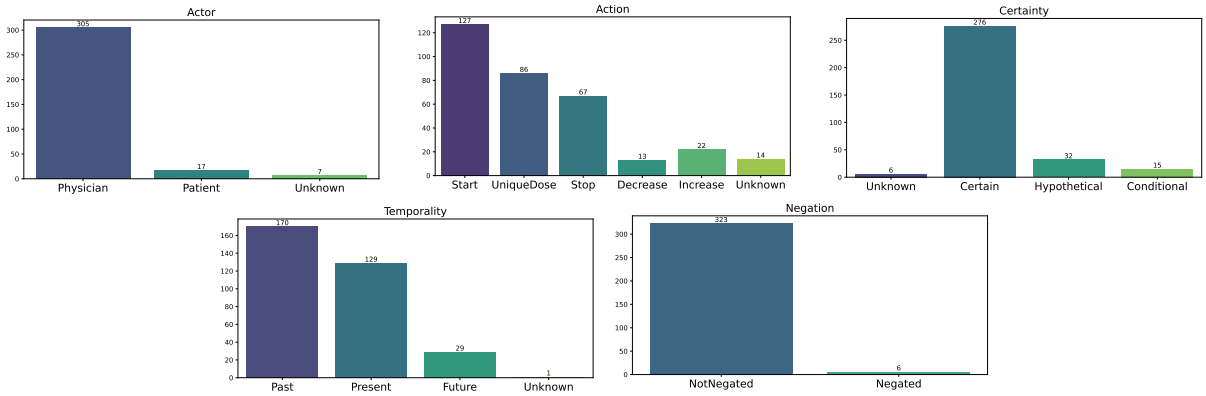


Figure 1: Distribution of context labels across CME2 trainset.

dataset with modified linguistic features. For instance, compared to the ENSEMBLE, when the model is trained with the dataset with modified #unique_nouns feature, the performance of the ENSEMBLE changes in an expected manner (See table 2) for Action, Actor, Temporality, highlighted by green, and unexpected manner for Certainty.

C Reproducibility

All the model were trained on the same hyperparameter setting.

Batch Size	4
Learning Rate	1e-5
Loss Function	Focal Loss
EPOCHS	20
Early stopping	True
Patience	3
MAX LEN	128

Table 10: Hyperparameter Setup.

D Linguistic Features

Table 19-20 list all the linguistic features that can be extracted from BEA (Lee and Lee, 2023) extractor tool.

Action	Actor	Certainty	Negation	Temporality
avg_brysbaert_age_of_acquisition_of_words (avgworddiff, lex-sem)	#unique_verbs (partofspeech, syntax)	corrected_verbs_variation (lexicalvariation, lex-sem)	simple_adpositions_variation (avgworddiff, lex-sem)	#unique_punctuations (partofspeech, syntax)
corrected_adverbs_variation (lexicalvariation, lex-sem)	#named_entities_cardinal (entity, discourse)	root_verbs_variation (lexicalvariation, lex-sem)	#named_entities_product (entity, discourse)	simple_type_token_ratio_no_lemma (typetokenratio, lex-sem)
root_adverbs_variation (avgworddiff, lex-sem)	avg_adverbs_per_word (partofspeech, syntax)	avg_particles_per_word (partofspeech, syntax)	smog_index (readformula, surface)	simple_type_token_ratio (typetokenratio, lex-sem)
#unique_nouns (partofspeech, syntax)	#named_entities_loc (entity, discourse)	avg_adjectives_per_word (partofspeech, syntax)	avg_named_entities_loc_per_word (entity, discourse)	simple_punctuation_variation (lexicalvariation, lex-sem)
simple_numerals_variation (lexicalvariation, lex-sem)	avg_named_entities_norp_per_sent (entity, discourse)	#unique_nouns (avgworddiff, lex-sem)	avg_named_entities_time_per_sent (entity, discourse)	#unique_nouns (partofspeech, syntax)

Table 11: Top 5 significant linguistic features for each context category in the order of decreasing effect size.

Action	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
average_brysbaert_age_of_acquistion_of_words_pe...	-1.332967	0.528185	-2.523674	1.388597e-02	-2.386399	-0.279535
corrected_adverbs_variation	5175.404330	2283.831956	2.266106	2.653914e-02	620.445122	9730.363538
root_adverbs_variation	-5201.900492	2296.742110	-2.264904	2.661663e-02	-9782.608191	-621.192792
total_number_of_unique_nouns	-13.631678	6.105775	-2.232588	2.877686e-02	-25.809262	-1.454095
simple_numerals_variation	-12.722929	5.826538	-2.183617	3.234429e-02	-24.343593	-1.102265
total_kuperman_age_of_acquistion_of_words	37.321799	18.411220	2.027123	4.645964e-02	0.601778	74.041819
average_kuperman_age_of_acquistion_of_words_per...	-36.646022	18.083067	-2.026538	4.652102e-02	-72.711562	-0.580481

Table 12: Action

Actor	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
total_number_of_unique_verbs	-6.603329e+00	8.827628e-01	-7.480298	1.650764e-10	-8.363944e+00	-4.842714e+00
total_number_of_named_entities_cardinal	1.423989e-08	2.126493e-09	6.696419	4.446148e-09	9.998731e-09	1.848104e-08
average_number_of_adverbs_per_word	1.320402e+00	2.292887e-01	5.758689	2.077356e-07	8.631005e-01	1.777704e+00
total_number_of_named_entities_loc	2.985107e-08	5.366359e-09	5.562631	4.546192e-07	1.914821e-08	4.055394e-08
average_number_of_named_entities_norp_per_sentence	3.550854e-08	6.615868e-09	5.367178	9.835470e-07	2.231361e-08	4.870347e-08
average_number_of_named_entities_ordinal_per_se...	-1.419000e-08	2.679636e-09	-5.295494	1.302100e-06	-1.953436e-08	-8.845632e-09
total_number_of_named_entities_date	7.829471e-08	1.498465e-08	5.224995	1.713522e-06	4.840878e-08	1.081806e-07
simple_verbs_variation	-3.609825e+00	7.089304e-01	-5.091932	2.866104e-06	-5.023743e+00	-2.195908e+00
total_number_of_unique_adverbs	7.335495e+00	1.457505e+00	5.032912	3.594641e-06	4.428592e+00	1.024240e+01
total_number_of_named_entities_money	6.785820e-08	1.353844e-08	5.012262	3.890109e-06	4.085664e-08	9.485977e-08
simple_coordinating_conjunctions_variation	9.094260e+00	1.823606e+00	4.986966	4.284578e-06	5.457193e+00	1.273133e+01
simple_determiners_variation	-3.627148e+00	7.319451e-01	-4.955492	4.830352e-06	-5.086966e+00	-2.167329e+00
average_number_of_named_entities_date_per_word	6.510582e-08	1.318313e-08	4.938572	5.151270e-06	3.881290e-08	9.139874e-08
simple_adverbs_variation	9.767410e+00	1.989599e+00	4.909235	5.757781e-06	5.799280e+00	1.373554e+01
total_number_of_named_entities_norp	-3.300574e-07	6.728979e-08	-4.905014	5.850624e-06	-4.642626e-07	-1.958521e-07
root_adverbs_variation	-2.730079e+03	5.568992e+02	-4.902286	5.911379e-06	-3.840780e+03	-1.619379e+03
average_number_of_named_entities_law_per_word	-8.650389e-09	1.764684e-09	-4.901949	5.918936e-06	-1.216994e-08	-5.130839e-09
corrected_adverbs_variation	2.712329e+03	5.537689e+02	4.897944	6.009383e-06	1.607872e+03	3.816786e+03
average_number_of_coordinating_conjunctions_per...	9.809528e-01	2.006043e-01	4.889989	6.193042e-06	5.808601e-01	1.381045e+00
average_number_of_syllables_per_word	2.316189e+02	4.836869e+01	4.788613	9.072603e-06	1.351506e+02	3.280872e+02
total_number_of_named_entities_ordinal	-2.596594e-08	5.438161e-09	-4.774765	9.555796e-06	-3.681201e-08	-1.511987e-08
average_number_of_auxiliaries_per_word	1.251121e+00	2.645758e-01	4.728781	1.134728e-05	7.234414e-01	1.778801e+00
average_number_of_pronouns_per_word	1.311673e+00	2.800749e-01	4.683294	1.343990e-05	7.530815e-01	1.870265e+00
average_number_of_nouns_per_word	2.257798e+00	4.860383e-01	4.645308	1.547154e-05	1.288425e+00	3.227171e+00
total_number_of_unique_nouns	6.716098e+00	1.480489e+00	4.536405	2.309635e-05	3.763356e+00	9.668840e+00
average_number_of_named_entities_event_per_word	2.472929e-08	5.466195e-09	4.524040	2.416458e-05	1.382730e-08	3.563127e-08
total_number_of_unique_determiners	-3.375619e+00	7.541151e-01	-4.476264	2.876153e-05	-4.879654e+00	-1.871584e+00
root_verbs_variation	1.868781e+03	4.193529e+02	4.456344	3.091963e-05	1.032408e+03	2.705154e+03
corrected_verbs_variation	-1.860470e+03	4.178374e+02	-4.452618	3.134043e-05	-2.693820e+03	-1.027120e+03
average_number_of_named_entities_language_per_word	4.124240e-08	9.411679e-09	4.382045	4.044268e-05	2.247140e-08	6.001340e-08
total_number_of_named_entities_language	-4.495325e-08	1.068470e-08	-4.207255	7.539254e-05	-6.626321e-08	-2.364329e-08
average_number_of_adpositions_per_word	1.292568e+00	3.124311e-01	4.137131	9.644439e-05	6.694440e-01	1.915692e+00
root_auxiliaries_variation	7.020887e+02	1.715549e+02	4.092502	1.126829e-04	3.599333e+02	1.044244e+03
average_number_of_named_entities_law_per_sentence	-1.571381e-08	3.840805e-09	-4.091279	1.131628e-04	-2.337405e-08	-8.053563e-09
coleman_liau_index	3.698407e+00	9.062749e-01	4.080889	1.173224e-04	1.890899e+00	5.505916e+00
corrected_auxiliaries_variation	-6.962670e+02	1.707454e+02	-4.077808	1.185835e-04	-1.036808e+03	-3.557260e+02
average_number_of_named_entities_org_per_sentence	1.349471e-08	3.326202e-09	4.057091	1.274113e-04	6.860804e-09	2.012861e-08
average_number_of_named_entities_date_per_sentence	1.978967e-08	4.912671e-09	4.028292	1.407396e-04	9.991661e-09	2.958769e-08
total_number_of_named_entities_time	7.983593e-08	2.011475e-08	3.969023	1.725146e-04	3.971832e-08	1.199535e-07
average_number_of_named_entities_fac_per_sentence	1.480192e-08	3.734692e-09	3.963358	1.758897e-04	7.353312e-09	2.225053e-08
average_number_of_named_entities_loc_per_sentence	1.433150e-08	3.616952e-09	3.962314	1.765185e-04	7.117716e-09	2.154528e-08
average_number_of_named_entities_language_per_s...	3.009464e-08	7.654272e-09	3.931745	1.959175e-04	1.482868e-08	4.536061e-08
average_number_of_spaces_per_word	1.032595e+00	2.675687e-01	3.859178	2.504993e-04	4.989463e-01	1.566244e+00
corrected_adjectives_variation	4.139136e+03	1.074507e+03	3.852126	2.565202e-04	1.996099e+03	6.282173e+03
root_adjectives_variation	-4.143990e+03	1.076025e+03	-3.851202	2.573188e-04	-6.290054e+03	-1.997926e+03
smog_index	-7.740051e-01	2.018465e-01	-3.834623	2.720674e-04	-1.176575e+00	-3.714350e-01
total_number_of_words_more_than_three_syllables	1.110019e+00	2.919974e-01	3.801468	3.040247e-04	5.276483e-01	1.692389e+00
average_number_of_verbs_per_word	1.098956e+00	2.953297e-01	3.721116	3.970383e-04	5.099396e-01	1.687973e+00
root_coordinating_conjunctions_variation	-5.918521e+03	1.630449e+03	-3.629995	5.353012e-04	-9.170348e+03	-2.666693e+03
corrected_coordinating_conjunctions_variation	5.907623e+03	1.629720e+03	3.624931	5.441962e-04	2.657249e+03	9.157997e+03
average_number_of_proper_nouns_per_word	1.291358e+00	3.613331e-01	3.573872	6.421061e-04	5.707021e-01	2.012015e+00
average_number_of_named_entities_quantity_per_s...	-6.742536e-09	1.893705e-09	-3.560499	6.703914e-04	-1.051941e-08	-2.965659e-09
average_number_of_subordinating_conjunctions_pe...	4.302424e-01	1.223792e-01	3.515651	7.741333e-04	1.861648e-01	6.743199e-01
average_number_of_named_entities_cardinal_per_s...	-2.051229e-08	5.878661e-09	-3.489279	8.420618e-04	-3.223691e-08	-8.787671e-09
average_number_of_named_entities_time_per_word	2.397853e-08	7.173665e-09	3.342577	1.335283e-03	9.671104e-09	3.828595e-08
reading_time_for_slow_readers	2.543545e+01	7.808767e+00	3.257294	1.736213e-03	9.861352e+00	4.100954e+01
average_number_of_named_entities_ordinal_per_word	-2.398542e-08	7.495725e-09	-3.199880	2.067104e-03	-3.893517e-08	-9.035665e-09
simple_nouns_variation	4.572277e-01	1.447439e-01	3.158873	2.338662e-03	1.685451e-01	7.459103e-01
total_number_of_unique_pronouns	-5.933595e+00	1.886614e+00	-3.145102	2.437108e-03	-9.696329e+00	-2.170861e+00
simple_pronouns_variation	-5.690443e+00	1.838407e+00	-3.095312	2.826292e-03	-9.357029e+00	-2.023856e+00
average_number_of_numerals_per_word	9.211165e-01	3.016055e-01	3.054044	3.192013e-03	3.195834e-01	1.522650e+00
average_number_of_adjectives_per_word	9.039703e-01	2.996039e-01	3.017218	3.555111e-03	3.064291e-01	1.501511e+00
average_number_of_named_entities_norp_per_word	3.013695e-08	1.038449e-08	2.902112	4.952210e-03	9.425735e-09	5.084816e-08
root_particles_variation	1.355049e+01	4.699383e+00	2.883462	5.221432e-03	4.177868e+00	2.292311e+01
corrected_particles_variation	1.369325e+01	4.750674e+00	2.882380	5.237457e-03	4.218327e+00	2.316817e+01
root_adpositions_variation	2.611851e+02	9.240457e+01	2.826539	6.128892e-03	7.689001e+01	4.454802e+02
total_number_of_named_entities_org	7.365727e-08	2.613092e-08	2.818778	6.263278e-03	2.154080e-08	1.257737e-07
corrected_adpositions_variation	-2.592618e+02	9.275231e+01	-2.795206	6.688263e-03	-4.442505e+02	-7.427314e+01
simple_particles_variation	-1.393290e+01	5.031651e+00	-2.769051	7.190673e-03	-2.396821e+01	-3.897588e+00
root_determiners_variation	8.658698e+02	3.227626e+02	2.682683	9.105436e-03	2.221401e+02	1.509599e+03
total_number_of_unique_numerals	4.208560e+00	1.576313e+00	2.669876	9.426012e-03	1.064703e+00	7.352417e+00

Table 13: Actor - Part1

Actor	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
corrected_determiners_variation	-8.581236e+02	3.219952e+02	-2.665020	9.550229e-03	-1.500323e+03	-2.159244e+02
average_number_of_named_entities_gpe_per_word	1.416567e-08	5.513924e-09	2.569072	1.233065e-02	3.168492e-09	2.516284e-08
total_number_of_unique_particles	-1.333033e+01	5.194925e+00	-2.566030	1.242971e-02	-2.369128e+01	-2.969382e+00
average_number_of_named_entities_gpe_per_sentence	-1.223607e-08	4.854813e-09	-2.520399	1.400475e-02	-2.191869e-08	-2.553447e-09
total_number_of_named_entities_fac	-1.487456e-07	6.189584e-08	-2.403160	1.890746e-02	-2.721930e-07	-2.529826e-08
total_number_of_named_entities_law	7.439520e-08	3.105679e-08	2.395456	1.927782e-02	1.245437e-08	1.363360e-07
simple_adjectives_variation	3.805237e+00	1.590028e+00	2.393189	1.938805e-02	6.340270e-01	6.976447e+00
total_number_of_named_entities_quantity	5.019254e-09	2.127752e-09	2.358947	2.112259e-02	7.755860e-10	9.262923e-09
average_number_of_named_entities_percent_per_se...	-4.051988e-09	1.740027e-09	-2.328693	2.276858e-02	-7.522361e-09	-5.816140e-10
average_number_of_named_entities_person_per_sen...	6.094086e-09	2.686051e-09	2.268790	2.636675e-02	7.369257e-10	1.145125e-08
root_type_token_ratio	-9.129534e+01	4.082464e+01	-2.236281	2.852244e-02	-1.727175e+02	-9.873173e+00
root_type_token_ratio_no_lemma	-9.129534e+01	4.082464e+01	-2.236281	2.852244e-02	-1.727175e+02	-9.873173e+00
corrected_type_token_ratio	8.965713e+01	4.088803e+01	2.192748	3.165116e-02	8.108534e+00	1.712057e+02
corrected_type_token_ratio_no_lemma	8.965713e+01	4.088803e+01	2.192748	3.165116e-02	8.108534e+00	1.712057e+02
total_number_of_named_entities_gpe	-1.436278e-08	6.634177e-09	-2.164968	3.380136e-02	-2.759423e-08	-1.131333e-09
simple_auxiliaries_variation	-2.849675e+00	1.345381e+00	-2.118117	3.771737e-02	-5.532954e+00	-1.663967e-01
average_number_of_named_entities_person_per_word	-2.567186e-08	1.213546e-08	-2.115442	3.795238e-02	-4.987526e-08	-1.468454e-09
average_number_of_named_entities_percent_per_word	1.101978e-08	5.301360e-09	2.078670	4.131448e-02	4.465514e-10	2.159301e-08
average_number_of_particles_per_word	3.744410e-01	1.805495e-01	2.073897	4.176934e-02	1.434636e-02	7.345356e-01
simple_adpositions_variation	-9.511057e-01	4.714977e-01	-2.017201	4.751080e-02	-1.891478e+00	-1.073309e-02
average_number_of_characters_per_word	-2.612976e+02	1.305870e+02	-2.000947	4.927703e-02	-5.217451e+02	-8.500675e-01

Table 14: Actor - Part2

Certainty	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
corrected_verbs_variation	-3.212384e+03	7.606418e+02	-4.223254	7.125228e-05	-4.729436e+03	-1.695331e+03
root_verbs_variation	3.221559e+03	7.634007e+02	4.220011	7.207333e-05	1.699005e+03	4.744114e+03
average_number_of_particles_per_word	1.256040e+00	3.286769e-01	3.821503	2.843090e-04	6.005144e-01	1.911565e+00
average_number_of_adjectives_per_word	1.975107e+00	5.454067e-01	3.621348	5.505765e-04	8.873279e-01	3.062887e+00
total_number_of_unique_nouns	9.644532e+00	2.695120e+00	3.578517	6.325497e-04	4.269285e+00	1.501978e+01
average_number_of_nouns_per_word	3.132926e+00	8.847966e-01	3.540843	7.141205e-04	1.368255e+00	4.897597e+00
average_number_of_spaces_per_word	1.704043e+00	4.870889e-01	3.498423	8.178939e-04	7.325748e-01	2.675511e+00
simple_verbs_variation	-4.506679e+00	1.290555e+00	-3.492047	8.346754e-04	-7.080610e+00	-1.932748e+00
average_number_of_coordinating_conjunctions_per...	1.272030e+00	3.651852e-01	3.483246	8.583776e-04	5.436909e-01	2.000369e+00
average_number_of_proper_nouns_per_word	2.275784e+00	6.577801e-01	3.459795	9.246828e-04	9.638833e-01	3.587685e+00
average_number_of_numerals_per_word	1.845811e+00	5.490503e-01	3.361825	1.257749e-03	7.507647e-01	2.940857e+00
total_number_of_words_more_than_three_syllables	1.747913e+00	5.315596e-01	3.288274	1.579013e-03	6.877512e-01	2.808075e+00
total_number_of_unique_verbs	-5.212617e+00	1.607004e+00	-3.243686	1.809805e-03	-8.417686e+00	-2.007549e+00
average_number_of_pronouns_per_word	1.622757e+00	5.098556e-01	3.182777	2.176561e-03	6.058818e-01	2.639632e+00
average_number_of_verbs_per_word	1.678832e+00	5.376258e-01	3.122677	2.605744e-03	6.065711e-01	2.751093e+00
total_number_of_named_entities_gpe	-3.761732e-08	1.207703e-08	-3.114784	2.667656e-03	-6.170419e-08	-1.353045e-08
average_number_of_adpositions_per_word	1.745591e+00	5.687575e-01	3.069130	3.053488e-03	6.112399e-01	2.879942e+00
smog_index	-1.118901e+00	3.674465e-01	-3.045073	3.277147e-03	-1.851750e+00	-3.860523e-01
simple_determiners_variation	-4.040180e+00	1.332452e+00	-3.032140	3.403587e-03	-6.697671e+00	-1.382689e+00
total_number_of_unique_numerals	8.620557e+00	2.869560e+00	3.004139	3.693049e-03	2.897399e+00	1.434371e+01
coleman_liau_index	4.932388e+00	1.649806e+00	2.989677	3.851343e-03	1.641953e+00	8.222822e+00
total_number_of_named_entities_quantity	-1.147253e-08	3.873415e-09	-2.961865	4.173613e-03	-1.919781e-08	-3.747248e-09
total_number_of_unique_adverbs	7.468378e+00	2.653280e+00	2.814772	6.333705e-03	2.176579e+00	1.276018e+01
bilogarithmic_type_token_ratio	-6.978865e-01	2.541982e-01	-2.745443	7.673647e-03	-1.204869e+00	-1.909043e-01
bilogarithmic_type_token_ratio_no_lemma	-6.978865e-01	2.541982e-01	-2.745443	7.673652e-03	-1.204869e+00	-1.909042e-01
total_kuperman_age_of_acquistion_of_words	-2.174240e+01	8.126806e+00	-2.675392	9.286694e-03	-3.795080e+01	-5.533992e+00
average_number_of_named_entities_cardinal_per_word	1.447710e-08	5.514504e-09	2.625278	1.062418e-02	3.478774e-09	2.547544e-08
simple_type_token_ratio	1.541694e+00	6.013573e-01	2.563691	1.250639e-02	3.423247e-01	2.741063e+00
simple_type_token_ratio_no_lemma	1.541694e+00	6.013573e-01	2.563691	1.250639e-02	3.423247e-01	2.741063e+00
average_kuperman_age_of_acquistion_of_words_per...	2.037079e+01	7.981958e+00	2.552104	1.289251e-02	4.451275e+00	3.629030e+01
simple_numerals_variation	6.483988e+00	2.571864e+00	2.521124	1.397837e-02	1.354567e+00	1.161341e+01
average_number_of_named_entities_product_per_se...	-1.031647e-08	4.109783e-09	-2.510223	1.437966e-02	-1.851318e-08	-2.119768e-09
total_number_of_unique_determiners	-3.325580e+00	1.372811e+00	-2.422461	1.800730e-02	-6.063565e+00	-5.875961e-01
average_number_of_adverbs_per_word	9.997150e-01	4.174031e-01	2.395083	1.929594e-02	1.672309e-01	1.832199e+00
total_number_of_named_entities_cardinal	-9.126049e-09	3.871122e-09	-2.357468	2.120049e-02	-1.684676e-08	-1.405338e-09
flesch_kincaid_grade_level	3.737562e+03	1.594155e+03	2.344541	2.189264e-02	5.581205e+02	6.917003e+03
flesch_kincaid_reading_ease	2.999031e+03	1.286925e+03	2.330386	2.267354e-02	4.323410e+02	5.565722e+03
reading_time_for_fast_readers	-1.356564e+01	5.852540e+00	-2.317907	2.338238e-02	-2.523817e+01	-1.893119e+00
average_number_of_named_entities_money_per_word	2.653509e-08	1.160470e-08	2.286581	2.524896e-02	3.390247e-09	4.967994e-08
root_type_token_ratio_no_lemma	-1.652804e+02	7.431822e+01	-2.223955	2.937941e-02	-3.135034e+02	-1.705738e+01
root_type_token_ratio	-1.652804e+02	7.431822e+01	-2.223955	2.937941e-02	-3.135034e+02	-1.705738e+01
total_number_of_unique_particles	2.098733e+01	9.456974e+00	2.219244	2.971291e-02	2.125994e+00	3.984867e+01
total_number_of_unique_auxiliaries	-5.758467e+00	2.625123e+00	-2.193599	3.158723e-02	-1.099411e+01	-5.228236e-01
average_number_of_subordinating_conjunctions_pe...	4.876051e-01	2.227822e-01	2.188708	3.195621e-02	4.328008e-02	9.319301e-01
total_number_of_unique_coordinating_conjunctions	4.736511e+00	2.166110e+00	2.186644	3.211308e-02	4.163406e-01	9.056682e+00
corrected_type_token_ratio_no_lemma	1.619719e+02	7.443361e+01	2.176059	3.292807e-02	1.351877e+01	3.104251e+02
corrected_type_token_ratio	1.619719e+02	7.443361e+01	2.176059	3.292807e-02	1.351877e+01	3.104251e+02
average_number_of_proper_nouns_per_sentence	-6.589614e+01	3.039285e+01	-2.168146	3.354909e-02	-1.265128e+02	-5.279508e+00
total_number_of_named_entities_ordinal	2.134446e-08	9.899768e-09	2.156057	3.451764e-02	1.599997e-09	4.108893e-08
total_number_of_unique_proper_nouns	3.251199e+01	1.523059e+01	2.134651	3.629253e-02	2.135541e+00	6.288845e+01
total_number_of_proper_nouns	3.251199e+01	1.523059e+01	2.134651	3.629253e-02	2.135541e+00	6.288845e+01
root_particles_variation	-1.801303e+01	8.554877e+00	-2.105586	3.882938e-02	-3.507519e+01	-9.508670e-01
corrected_particles_variation	-1.820879e+01	8.648248e+00	-2.105489	3.883811e-02	-3.545718e+01	-9.604052e-01
root_determiners_variation	1.223518e+03	5.875652e+02	2.082353	4.096658e-02	5.165614e+01	2.395380e+03
average_number_of_named_entities_law_per_word	6.668953e-09	3.212475e-09	2.075955	4.157272e-02	2.618729e-10	1.307603e-08
corrected_determiners_variation	-1.214813e+03	5.861683e+02	-2.072465	4.190660e-02	-2.383889e+03	-4.573735e+01
average_number_of_verbs_per_sentence	-5.335525e+01	2.651491e+01	-2.012273	4.804043e-02	-1.062376e+02	-4.729229e-01
average_number_of_auxiliaries_per_sentence	-6.196236e+01	3.080374e+01	-2.011520	4.812178e-02	-1.233985e+02	-5.262309e-01
average_number_of_adpositions_per_sentence	-8.579195e+01	4.280723e+01	-2.004146	4.892499e-02	-1.711683e+02	-4.156157e-01
automated_readability_index	-6.747181e+02	3.371693e+02	-2.001125	4.925731e-02	-1.347181e+03	-2.255082e+00

Table 15: Certainty

Negation	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
simple_adpositions_variation	-1.719098e+00	6.155434e-01	-2.792814	6.732834e-03	-2.946761e+00	-4.914355e-01
total_number_of_named_entities_product	4.244266e-08	1.733488e-08	2.448397	1.685814e-02	7.869341e-09	7.701598e-08
smog_index	-6.244617e-01	2.635119e-01	-2.369767	2.056014e-02	-1.150020e+00	-9.890378e-02
average_number_of_named_entities_loc_per_word	5.714464e-08	2.488083e-08	2.296733	2.463010e-02	7.521382e-09	1.067679e-07
average_number_of_named_entities_time_per_sentence	-1.229495e-08	5.436076e-09	-2.261733	2.682211e-02	-2.313686e-08	-1.453039e-09
total_number_of_named_entities_percent	9.312301e-08	4.136844e-08	2.251064	2.752364e-02	1.061625e-08	1.756298e-07
total_number_of_named_entities_law	9.120490e-08	4.054485e-08	2.249482	2.762905e-02	1.034075e-08	1.720690e-07
average_number_of_named_entities_cardinal_per_s...	-1.719405e-08	7.674631e-09	-2.240375	2.824268e-02	-3.250062e-08	-1.887482e-09
average_number_of_named_entities_gpe_per_sentence	-1.393150e-08	6.337991e-09	-2.198094	3.125139e-02	-2.657222e-08	-1.290775e-09
average_number_of_named_entities_org_per_word	4.446542e-08	2.024354e-08	2.196524	3.136830e-02	4.090961e-09	8.483988e-08
average_number_of_named_entities_money_per_sent...	2.024129e-08	9.350754e-09	2.164669	3.382518e-02	1.591797e-09	3.889078e-08
total_number_of_named_entities_fac	-1.743072e-07	8.080543e-08	-2.157123	3.443124e-02	-3.354686e-07	-1.314590e-08
total_number_of_named_entities_event	6.852808e-08	3.183012e-08	2.152932	3.477191e-02	5.044914e-09	1.320112e-07
average_number_of_named_entities_art_per_sentence	2.453337e-08	1.147241e-08	2.138468	3.597032e-02	1.652384e-09	4.741436e-08
average_number_of_named_entities_loc_per_sentence	1.000902e-08	4.721955e-09	2.119678	3.758082e-02	5.913809e-10	1.942667e-08
average_number_of_named_entities_gpe_per_word	1.519277e-08	7.198464e-09	2.110557	3.838489e-02	8.358857e-10	2.954965e-08
average_number_of_named_entities_art_per_word	-5.169465e-08	2.462123e-08	-2.099597	3.937085e-02	-1.008001e-07	-2.589165e-09
simple_coordinating_conjunctions_variation	4.958396e+00	2.380729e+00	2.082721	4.093188e-02	2.101811e-01	9.706611e+00
average_number_of_named_entities_percent_per_word	1.438223e-08	6.920960e-09	2.078069	4.137159e-02	5.788102e-10	2.818565e-08
average_number_of_named_entities_fac_per_word	-3.128243e-08	1.506596e-08	-2.076365	4.153365e-02	-6.133055e-08	-1.234316e-09
root_coordinating_conjunctions_variation	-4.413945e+03	2.128561e+03	-2.073675	4.179054e-02	-8.659227e+03	-1.686630e+02
corrected_coordinating_conjunctions_variation	4.409099e+03	2.127610e+03	2.072325	4.192006e-02	1.657140e+02	8.652483e+03
average_number_of_named_entities_event_per_sent...	1.840485e-08	8.899373e-09	2.068107	4.232679e-02	6.556136e-10	3.615409e-08
average_number_of_named_entities_person_per_word	-3.237656e-08	1.584292e-08	-2.043598	4.475826e-02	-6.397426e-08	-7.788546e-10
average_number_of_named_entities_language_per_word	2.490433e-08	1.228701e-08	2.026883	4.648480e-02	3.986602e-10	4.941000e-08
total_number_of_named_entities_person	8.564630e-08	4.238720e-08	2.020570	4.715160e-02	1.107697e-09	1.701849e-07

Table 16: Negation

Temporality	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
total_number_of_unique_punctuations	-8.488106e+00	1.458012e+00	-5.821699	1.612211e-07	-1.139602e+01	-5.580193e+00
simple_type_token_ratio_no_lemma	4.004287e+00	6.977944e-01	5.738491	2.252798e-07	2.612580e+00	5.395994e+00
simple_type_token_ratio	4.004287e+00	6.977943e-01	5.738491	2.252798e-07	2.612580e+00	5.395994e+00
simple_punctuations_variation	-5.202325e+00	9.263226e-01	-5.616105	3.674957e-07	-7.049817e+00	-3.354833e+00
total_number_of_unique_nouns	1.714003e+01	3.127325e+00	5.480731	6.289046e-07	1.090277e+01	2.337728e+01
total_number_of_unique_particles	5.991561e+01	1.097355e+01	5.460003	6.825611e-07	3.802956e+01	8.180166e+01
simple_auxiliaries_variation	-1.550611e+01	2.841929e+00	-5.456191	6.929087e-07	-2.117416e+01	-9.838059e+00
corrected_particles_variation	-5.234579e+01	1.003513e+01	-5.216255	1.772690e-06	-7.236023e+01	-3.233136e+01
root_particles_variation	-5.177689e+01	9.926785e+00	-5.215876	1.775294e-06	-7.157523e+01	-3.197854e+01
average_number_of_punctuations_per_word	-3.707239e+00	7.182851e-01	-5.161235	2.193892e-06	-5.139813e+00	-2.274664e+00
simple_particles_variation	5.464536e+01	1.062865e+01	5.141324	2.369345e-06	3.344717e+01	7.584354e+01
total_number_of_unique_auxiliaries	-1.535464e+01	3.046103e+00	-5.040749	3.488348e-06	-2.142990e+01	-9.279380e+00
coleman_liau_index	9.580976e+00	1.914378e+00	5.004745	4.003460e-06	5.762868e+00	1.339908e+01
average_number_of_adpositions_per_word	3.269788e+00	6.599667e-01	4.954474	4.849078e-06	1.953526e+00	4.586050e+00
total_number_of_words_more_than_three_syllables	2.960606e+00	6.168035e-01	4.799918	8.695859e-06	1.730431e+00	4.190782e+00
simple_verbs_variation	-7.145826e+00	1.497516e+00	-4.771787	9.662941e-06	-1.013253e+01	-4.159125e+00
average_number_of_named_entities_quantity_per_word	-4.121338e-08	8.651680e-09	-4.763627	9.962572e-06	-5.846861e-08	-2.395814e-08
average_number_of_interjections_per_word	1.792648e+00	3.835171e-01	4.674232	1.389958e-05	1.027747e+00	2.557549e+00
corrected_nouns_variation	-9.959689e+02	2.159900e+02	-4.611182	1.754958e-05	-1.426747e+03	-5.651905e+02
root_nouns_variation	9.975494e+02	2.166140e+02	4.605194	1.794119e-05	5.655264e+02	1.429572e+03
simple_proper_nouns_variation	9.242736e+00	2.047195e+00	4.514829	2.499143e-05	5.159734e+00	1.332574e+01
average_number_of_named_entities_art_per_word	1.797830e-07	3.983808e-08	4.512842	2.517333e-05	1.003284e-07	2.592375e-07
average_number_of_named_entities_percent_per_word	-5.033151e-08	1.119838e-08	-4.494536	2.691110e-05	-7.266598e-08	-2.799705e-08
total_number_of_unique_adpositions	1.096727e+01	2.450401e+00	4.475704	2.882018e-05	6.080100e+00	1.585444e+01
total_number_of_words_more_than_two_syllables	4.532242e+00	1.019734e+00	4.444535	3.227238e-05	2.498447e+00	6.566036e+00
root_proper_nouns_variation	-3.887525e+03	8.832845e+02	-4.401215	3.774369e-05	-5.649180e+03	-2.125870e+03
corrected_proper_nouns_variation	3.870173e+03	8.797788e+02	4.399030	3.804234e-05	2.115510e+03	5.624837e+03
root_verbs_variation	3.864364e+03	8.858240e+02	4.362452	4.339410e-05	2.097644e+03	5.631085e+03
corrected_verbs_variation	-3.847501e+03	8.826227e+02	-4.359168	4.390870e-05	-5.607836e+03	-2.087165e+03
average_number_of_named_entities_fac_per_word	1.061814e-07	2.437730e-08	4.355750	4.445068e-05	5.756244e-08	1.548004e-07
root_punctuations_variation	7.581011e+02	1.749593e+02	4.333015	4.822404e-05	4.091559e+02	1.107046e+03
total_number_of_named_entities_event	-2.224247e-07	5.150234e-08	-4.318730	5.075161e-05	-3.251429e-07	-1.197066e-07
corrected_punctuations_variation	-7.458231e+02	1.732774e+02	-4.304215	5.345123e-05	-1.091414e+03	-4.002322e+02
total_number_of_named_entities_product	-1.185646e-07	2.804849e-08	-4.227129	7.028313e-05	-1.745055e-07	-6.262363e-08
average_number_of_named_entities_event_per_sent...	-6.009121e-08	1.439952e-08	-4.173139	8.501091e-05	-8.881016e-08	-3.137227e-08
average_subtlex_us_zipf_of_words_per_word	-3.918426e+00	9.524894e-01	-4.113878	1.046008e-04	-5.818106e+00	-2.018745e+00
average_number_of_numerals_per_word	2.587804e+00	6.370992e-01	4.061854	1.253275e-04	1.317149e+00	3.858458e+00
average_number_of_auxiliaries_per_word	2.262518e+00	5.588793e-01	4.048313	1.313395e-04	1.147869e+00	3.377168e+00
gunning_fog_index	-2.912632e+00	7.260163e-01	-4.011800	1.489660e-04	-4.360626e+00	-1.464638e+00
simple_nouns_variation	-1.215635e+00	3.057511e-01	-3.975896	1.685040e-04	-1.825436e+00	-6.058333e-01
total_kuperman_age_of_acquistion_of_words	-3.737562e+01	9.430067e+00	-3.963452	1.758329e-04	-5.618330e+01	-1.856795e+01
total_number_of_unique_spaces	5.749141e+01	1.456223e+01	3.947981	1.853725e-04	2.844796e+01	8.653487e+01
total_number_of_named_entities_percent	-2.633546e-07	6.693571e-08	-3.934441	1.941272e-04	-3.968537e-07	-1.298556e-07
average_number_of_named_entities_org_per_word	-1.288588e-07	3.275481e-08	-3.934041	1.943917e-04	-1.941862e-07	-6.353138e-08
average_number_of_named_entities_loc_per_word	-1.580689e-07	4.025814e-08	-3.926385	1.995235e-04	-2.383613e-07	-7.777661e-08
root_adjectives_variation	-8.902584e+03	2.272952e+03	-3.916750	2.061655e-04	-1.343584e+04	-4.369325e+03
corrected_adjectives_variation	8.889159e+03	2.269745e+03	3.916369	2.064328e-04	4.362295e+03	1.341602e+04
average_number_of_nouns_per_word	4.013621e+00	1.026688e+00	3.909291	2.114528e-04	1.965957e+00	6.061285e+00
average_number_of_named_entities_time_per_sentence	3.425673e-08	8.795779e-09	3.894679	2.221910e-04	1.671410e-08	5.179936e-08
corrected_spaces_variation	-5.365895e+01	1.386591e+01	-3.869848	2.416467e-04	-8.131363e+01	-2.600427e+01
simple_spaces_variation	5.451684e+01	1.428988e+01	3.815067	2.905048e-04	2.601658e+01	8.301710e+01
average_kuperman_age_of_acquistion_of_words_per...	3.524573e+01	9.261990e+00	3.805416	3.000387e-04	1.677327e+01	5.371818e+01
reading_time_for_fast_readers	-2.574417e+01	6.791087e+00	-3.790877	3.149690e-04	-3.928857e+01	-1.219978e+01
root_spaces_variation	-4.460722e+01	1.180521e+01	-3.778605	3.281203e-04	-6.815197e+01	-2.106248e+01
average_number_of_named_entities_money_per_sent...	-5.695573e-08	1.512988e-08	-3.764454	3.439378e-04	-8.713132e-08	-2.678014e-08
average_number_of_named_entities_person_per_word	9.622665e-08	2.563444e-08	3.753803	3.563216e-04	4.510036e-08	1.473529e-07
total_brysbaert_age_of_acquistion_of_words	2.551302e+01	6.819914e+00	3.740959	3.718210e-04	1.191113e+01	3.911491e+01
total_number_of_named_entities_law	-2.418963e-07	6.560311e-08	-3.687269	4.438607e-04	-3.727376e-07	-1.110550e-07
average_number_of_named_entities_art_per_sentence	-6.825043e-08	1.856279e-08	-3.676733	4.594789e-04	-1.052727e-07	-3.122811e-08
total_number_of_unique_words	-2.855957e+01	7.842706e+00	-3.641545	5.155262e-04	-4.420135e+01	-1.291778e+01
total_number_of_named_entities_fac	4.710148e-07	1.307463e-07	3.602511	5.852995e-04	2.102496e-07	7.317800e-07
average_number_of_named_entities_percent_per_se...	-1.321096e-08	3.675562e-09	-3.594270	6.011350e-04	-2.054164e-08	-5.880286e-09
total_number_of_named_entities_art	1.704759e-07	4.760417e-08	3.581112	6.272691e-04	7.553235e-08	2.654194e-07
average_brysbaert_age_of_acquistion_of_words_pe...	-2.492547e+01	7.055126e+00	-3.532959	7.324102e-04	-3.899647e+01	-1.085447e+01
average_number_of_named_entities_product_per_word	9.089588e-08	2.590158e-08	3.509279	7.900521e-04	3.923681e-08	1.425550e-07
total_number_of_sentences	2.287950e+01	6.519742e+00	3.509265	7.900892e-04	9.876285e+00	3.588272e+01
average_number_of_verbs_per_word	2.158172e+00	6.238425e-01	3.459483	9.255977e-04	9.139578e-01	3.402387e+00
average_number_of_named_entities_cardinal_per_s...	4.226246e-08	1.241785e-08	3.403365	1.104648e-03	1.749585e-08	6.702907e-08
simple_adpositions_variation	3.327248e+00	9.959727e-01	3.340702	1.343072e-03	1.340843e+00	5.313653e+00
average_number_of_named_entities_person_per_sen...	1.894099e-08	5.673906e-09	3.338263	1.353269e-03	7.624741e-09	3.025724e-08
average_number_of_particles_per_word	1.264886e+00	3.813854e-01	3.316555	1.447285e-03	5.042365e-01	2.025535e+00

Table 17: Temporality-Part1

Temporality	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
simple_pronouns_variation	-1.276292e+01	3.883376e+00	-3.286553	1.587379e-03	-2.050807e+01	-5.017773e+00
average_kuperman_age_of_acquistion_of_words_per...	9.428420e-01	2.882325e-01	3.271116	1.664341e-03	3.679803e-01	1.517704e+00
average_number_of_named_entities_language_per_word	-6.366250e-08	1.988085e-08	-3.202202	2.052643e-03	-1.033136e-07	-2.401140e-08
bilogarithmic_type_token_ratio	-9.381622e-01	2.949628e-01	-3.180611	2.190802e-03	-1.526447e+00	-3.498773e-01
bilogarithmic_type_token_ratio_no_lemma	-9.381621e-01	2.949628e-01	-3.180611	2.190803e-03	-1.526447e+00	-3.498773e-01
average_number_of_named_entities_gpe_per_sentence	3.259444e-08	1.025511e-08	3.178361	2.205693e-03	1.214127e-08	5.304762e-08
average_number_of_named_entities_norp_per_word	6.964400e-08	2.193577e-08	3.174905	2.228745e-03	2.589448e-08	1.133935e-07
average_number_of_characters_per_sentence	-1.453636e+02	4.604424e+01	-3.157042	2.351536e-03	-2.371960e+02	-5.353126e+01
total_number_of_named_entities	-2.994569e+01	9.508760e+00	-3.149274	2.406884e-03	-4.891031e+01	-1.098107e+01
smog_index	-1.334089e+00	4.263723e-01	-3.128930	2.557660e-03	-2.184462e+00	-4.837163e-01
total_number_of_characters	1.355410e+02	4.332840e+01	3.128224	2.563044e-03	4.912519e+01	2.219567e+02
average_number_of_stop_words_per_word	2.581518e+00	8.283695e-01	3.116384	2.654992e-03	9.293867e-01	4.233648e+00
total_number_of_named_entities_gpe	-4.354429e-08	1.401377e-08	-3.107251	2.728013e-03	-7.149386e-08	-1.559471e-08
average_number_of_named_entities_per_sentence	2.983344e+01	9.603292e+00	3.106585	2.733412e-03	1.068028e+01	4.898660e+01
root_coordinating_conjunctions_variation	1.058921e+04	3.444094e+03	3.074600	3.004668e-03	3.720180e+03	1.745824e+04
average_number_of_named_entities_quantity_per_s...	-1.229481e-08	4.000186e-09	-3.073559	3.013899e-03	-2.027293e-08	-4.316689e-09
corrected_coordinating_conjunctions_variation	-1.057527e+04	3.442554e+03	-3.071925	3.028455e-03	-1.744123e+04	-3.709309e+03
simple_coordinating_conjunctions_variation	-1.181247e+01	3.852111e+00	-3.066492	3.077308e-03	-1.949526e+01	-4.129673e+00
corrected_adpositions_variation	-5.967113e+02	1.959262e+02	-3.045592	3.272164e-03	-9.874738e+02	-2.059487e+02
average_number_of_spaces_per_word	1.706844e+00	5.652012e-01	3.019887	3.527553e-03	5.795856e-01	2.834102e+00
root_adpositions_variation	5.875629e+02	1.951917e+02	3.010184	3.628681e-03	1.982654e+02	9.768603e+02
root_pronouns_variation	1.358380e+03	4.603614e+02	2.950682	4.310101e-03	4.402180e+02	2.276542e+03
corrected_pronouns_variation	-1.332799e+03	4.538840e+02	-2.936432	4.490009e-03	-2.238043e+03	-4.275563e+02
average_number_of_named_entities_time_per_word	-4.445738e-08	1.515336e-08	-2.933830	4.523594e-03	-7.467980e-08	-1.423496e-08
average_number_of_stop_words_per_sentence	-1.660031e+02	5.700572e+01	-2.912043	4.814133e-03	-2.796974e+02	-5.230877e+01
average_number_of_syllables_per_word	-2.972141e+02	1.021721e+02	-2.908957	4.856657e-03	-5.009898e+02	-9.343833e+01
uber_type_token_ratio	-3.692130e-01	1.272081e-01	-2.902434	4.947675e-03	-6.229215e-01	-1.155045e-01
uber_type_token_ratio_no_lemma	-3.692127e-01	1.272080e-01	-2.902432	4.947696e-03	-6.229211e-01	-1.155043e-01
average_number_of_subordinating_conjunctions_pe...	7.495760e-01	2.585088e-01	2.899615	4.987489e-03	2.339965e-01	1.265155e+00
average_number_of_named_entities_per_word	8.542054e-01	2.951073e-01	2.894558	5.059662e-03	2.656324e-01	1.442778e+00
average_number_of_interjections_per_sentence	-6.440576e+00	2.232746e+00	-2.884599	5.204637e-03	-1.089365e+01	-1.987505e+00
total_number_of_puntuations	3.316925e+01	1.154945e+01	2.871933	5.394545e-03	1.013460e+01	5.620390e+01
total_number_of_punctuations	3.316925e+01	1.154945e+01	2.871933	5.394546e-03	1.013460e+01	5.620390e+01
average_number_of_subordinating_conjunctions_pe...	-1.982583e+01	6.934693e+00	-2.858934	5.596079e-03	-3.365664e+01	-5.995022e+00
simple_adjectives_variation	9.458067e+00	3.358710e+00	2.815982	6.312358e-03	2.759332e+00	1.615680e+01
corrected_numerals_variation	2.361105e+03	8.449339e+02	2.794426	6.702768e-03	6.759375e+02	4.046273e+03
root_numerals_variation	-2.369576e+03	8.488371e+02	-2.791555	6.756402e-03	-4.062528e+03	-6.766232e+02
average_number_of_named_entities_gpe_per_word	-3.240184e-08	1.164739e-08	-2.781897	6.939742e-03	-5.563182e-08	-9.171852e-09
total_number_of_stop_words	1.489059e+02	5.454004e+01	2.730213	8.000787e-03	4.012924e+01	2.576826e+02
total_number_of_named_entities_person	-1.822952e-07	6.858410e-08	-2.657981	9.732926e-03	-3.190819e-07	-4.550856e-08
average_number_of_punctuations_per_sentence	-5.759106e+01	2.174557e+01	-2.648404	9.986602e-03	-1.009612e+02	-1.422089e+01
total_number_of_named_entities_quantity	-1.189930e-08	4.494578e-09	-2.647478	1.001144e-02	-2.086345e-08	-2.935145e-09
simple_adverbs_variation	-1.106562e+01	4.202748e+00	-2.632950	1.040869e-02	-1.944774e+01	-2.683507e+00
average_number_of_named_entities_date_per_sentence	-2.714601e-08	1.037733e-08	-2.615896	1.089321e-02	-4.784293e-08	-6.449081e-09
average_number_of_named_entities_loc_per_sentence	-1.948994e-08	7.640303e-09	-2.550938	1.293195e-02	-3.472805e-08	-4.251837e-09
total_number_of_unique_adjectives	5.061019e+01	1.986889e+01	2.547207	1.305892e-02	1.098293e+01	9.023744e+01
total_number_of_spaces	-3.129192e+01	1.242049e+01	-2.519378	1.404195e-02	-5.606381e+01	-6.520029e+00
average_number_of_named_entities_language_per_s...	-4.039236e-08	1.616857e-08	-2.498202	1.483419e-02	-7.263956e-08	-8.145154e-09
root_adverbs_variation	2.935009e+03	1.176371e+03	2.494968	1.495863e-02	5.888104e+02	5.281208e+03
total_number_of_unique_pronouns	-9.932665e+00	3.985208e+00	-2.492383	1.505879e-02	-1.788091e+01	-1.984419e+00
corrected_adverbs_variation	-2.913601e+03	1.169759e+03	-2.490770	1.512159e-02	-5.246611e+03	-5.805901e+02
total_number_of_adverbs	-4.744528e+01	1.920919e+01	-2.469926	1.595468e-02	-8.575681e+01	-9.133750e+00
total_number_of_named_entities_loc	-2.761345e-08	1.133568e-08	-2.435977	1.739997e-02	-5.022175e-08	-5.005156e-09
total_number_of_unique_determiners	3.816127e+00	1.592962e+00	2.395617	1.927005e-02	6.390641e-01	6.993190e+00
total_number_of_subordinating_conjunctions	1.573261e+01	6.600784e+00	2.383445	1.986821e-02	2.567758e+00	2.889746e+01
total_subtlex_us_zipf_of_words	5.537261e+01	2.323442e+01	2.383214	1.987971e-02	9.033015e+00	1.017122e+02
total_number_of_syllables	-1.370393e+02	5.794875e+01	-2.364836	2.081476e-02	-2.526144e+02	-2.146416e+01
average_number_of_named_entities_ordinal_per_word	3.711067e-08	1.583367e-08	2.343783	2.193388e-02	5.531423e-09	6.868993e-08
automated_readability_index	-9.018830e+02	3.912397e+02	-2.305193	2.412469e-02	-1.682186e+03	-1.215800e+02
total_number_of_named_entities_date	-7.290189e-08	3.165296e-08	-2.303162	2.424519e-02	-1.360317e-07	-9.772050e-09
average_number_of_characters_per_word	6.271808e+02	2.758466e+02	2.273658	2.605667e-02	7.702202e+01	1.177340e+03
average_number_of_adjectives_per_word	1.435990e+00	6.328712e-01	2.269008	2.635280e-02	1.737679e-01	2.698212e+00
total_number_of_named_entities_language	-5.113484e-08	2.256992e-08	-2.265619	2.657052e-02	-9.614914e-08	-6.120547e-09
total_number_of_pronouns	-1.088058e+02	4.855724e+01	-2.240773	2.821561e-02	-2.056501e+02	-1.196139e+01
total_number_of_determiners	-1.064012e+02	4.761202e+01	-2.234755	2.862733e-02	-2.013604e+02	-1.144201e+01
total_number_of_named_entities_time	-9.358431e-08	4.248959e-08	-2.202523	3.092355e-02	-1.783271e-07	-8.841489e-09
average_number_of_named_entities_event_per_word	2.519426e-08	1.154657e-08	2.181970	3.247077e-02	2.165356e-09	4.822316e-08
average_number_of_adverbs_per_word	1.048742e+00	4.843402e-01	2.165301	3.377489e-02	8.275605e-02	2.014728e+00
average_number_of_pronouns_per_sentence	9.992538e+01	4.657076e+01	2.145668	3.536939e-02	7.042922e+00	1.928078e+02
average_number_of_syllables_per_sentence	1.298597e+02	6.068938e+01	2.139743	3.586327e-02	8.818525e+00	2.509008e+02
average_number_of_determiners_per_sentence	1.010542e+02	4.728980e+01	2.136913	3.610127e-02	6.737672e+00	1.953707e+02
simple_determiners_variation	3.261927e+00	1.546131e+00	2.109735	3.845809e-02	1.782658e-01	6.345588e+00
average_number_of_coordinating_conjunctions_per...	5.408573e+01	2.591138e+01	2.087335	4.049987e-02	2.407114e+00	1.057643e+02
total_number_of_nouns	-1.968856e+02	9.514501e+01	-2.069322	4.220928e-02	-3.866464e+02	-7.124892e+00
total_number_of_adjectives	-1.159769e+02	5.607170e+01	-2.068368	4.230152e-02	-2.278084e+02	-4.145414e+00
total_number_of_coordinating_conjunctions	-5.389970e+01	2.619666e+01	-2.057503	4.336443e-02	-1.061473e+02	-1.652108e+00
average_number_of_pronouns_per_word	1.197110e+00	5.916190e-01	2.023447	4.684669e-02	1.716281e-02	2.377056e+00

Table 18: Temporality-Part2

	Key	Feature-name
0	t_word	total_number_of_words
1	t_stopword	total_number_of_stop_words
2	t_punct	total_number_of_puntuations
3	t_syll	total_number_of_syllables
4	t_syll2	total_number_of_words_more_than_two_syllables
5	t_syll3	total_number_of_words_more_than_three_syllables
6	t_uword	total_number_of_unique_words
7	t_sent	total_number_of_sentences
8	t_char	total_number_of_characters
9	a_word_ps	average_number_of_words_per_sentence
10	a_char_ps	average_number_of_characters_per_sentence
11	a_char_pw	average_number_of_characters_per_word
12	a_syll_ps	average_number_of_syllables_per_sentence
13	a_syll_pw	average_number_of_syllables_per_word
14	a_stopword_ps	average_number_of_stop_words_per_sentence
15	a_stopword_pw	average_number_of_stop_words_per_word
16	t_kup	total_kuperman_age_of_acquistion_of_words
17	t_bry	total_brysbaert_age_of_acquistion_of_words
18	t_subtlex_us_zipf	total_subtlex_us_zipf_of_words
19	a_kup_pw	average_kuperman_age_of_acquistion_of_words_pe...
20	a_bry_pw	average_brysbaert_age_of_acquistion_of_words_p...
21	a_kup_ps	average_kuperman_age_of_acquistion_of_words_pe...
22	a_bry_ps	average_brysbaert_age_of_acquistion_of_words_p...
23	a_subtlex_us_zipf_pw	average_subtlex_us_zipf_of_words_per_word
24	a_subtlex_us_zipf_ps	average_subtlex_us_zipf_of_words_per_sentence
25	t_n_ent	total_number_of_named_entities
26	t_n_ent_person	total_number_of_named_entities_person
27	t_n_ent_norp	total_number_of_named_entities_norp
28	t_n_ent_fac	total_number_of_named_entities_fac
29	t_n_ent_org	total_number_of_named_entities_org
30	t_n_ent_gpe	total_number_of_named_entities_gpe
31	t_n_ent_loc	total_number_of_named_entities_loc
32	t_n_ent_product	total_number_of_named_entities_product
33	t_n_ent_event	total_number_of_named_entities_event
34	t_n_ent_art	total_number_of_named_entities_art
35	t_n_ent_law	total_number_of_named_entities_law
36	t_n_ent_language	total_number_of_named_entities_language
37	t_n_ent_date	total_number_of_named_entities_date
38	t_n_ent_time	total_number_of_named_entities_time
39	t_n_ent_percent	total_number_of_named_entities_percent
40	t_n_ent_money	total_number_of_named_entities_money
41	t_n_ent_quantity	total_number_of_named_entities_quantity
42	t_n_ent_ordinal	total_number_of_named_entities_ordinal
43	t_n_ent_cardinal	total_number_of_named_entities_cardinal
44	a_n_ent_pw	average_number_of_named_entities_per_word
45	a_n_ent_person_pw	average_number_of_named_entities_person_per_word
46	a_n_ent_norp_pw	average_number_of_named_entities_norp_per_word
47	a_n_ent_fac_pw	average_number_of_named_entities_fac_per_word
48	a_n_ent_org_pw	average_number_of_named_entities_org_per_word
49	a_n_ent_gpe_pw	average_number_of_named_entities_gpe_per_word
50	a_n_ent_loc_pw	average_number_of_named_entities_loc_per_word
51	a_n_ent_product_pw	average_number_of_named_entities_product_per_word
52	a_n_ent_event_pw	average_number_of_named_entities_event_per_word
53	a_n_ent_art_pw	average_number_of_named_entities_art_per_word
54	a_n_ent_law_pw	average_number_of_named_entities_law_per_word

	Key	Feature-name
55	a_n_ent_language_pw	average_number_of_named_entities_language_per_...
56	a_n_ent_date_pw	average_number_of_named_entities_date_per_word
57	a_n_ent_time_pw	average_number_of_named_entities_time_per_word
58	a_n_ent_percent_pw	average_number_of_named_entities_percent_per_word
59	a_n_ent_money_pw	average_number_of_named_entities_money_per_word
60	a_n_ent_quantity_pw	average_number_of_named_entities_quantity_per_...
61	a_n_ent_ordinal_pw	average_number_of_named_entities_ordinal_per_word
62	a_n_ent_cardinal_pw	average_number_of_named_entities_cardinal_per_...
63	a_n_ent_ps	average_number_of_named_entities_per_sentence
64	a_n_ent_person_ps	average_number_of_named_entities_person_per_se...
65	a_n_ent_norp_ps	average_number_of_named_entities_norp_per_sent...
66	a_n_ent_fac_ps	average_number_of_named_entities_fac_per_sentence
67	a_n_ent_org_ps	average_number_of_named_entities_org_per_sentence
68	a_n_ent_gpe_ps	average_number_of_named_entities_gpe_per_sentence
69	a_n_ent_loc_ps	average_number_of_named_entities_loc_per_sentence
70	a_n_ent_product_ps	average_number_of_named_entities_product_per_s...
71	a_n_ent_event_ps	average_number_of_named_entities_event_per_sen...
72	a_n_ent_art_ps	average_number_of_named_entities_art_per_sentence
73	a_n_ent_law_ps	average_number_of_named_entities_law_per_sentence
74	a_n_ent_language_ps	average_number_of_named_entities_language_per_...
75	a_n_ent_date_ps	average_number_of_named_entities_date_per_sent...
76	a_n_ent_time_ps	average_number_of_named_entities_time_per_sent...
77	a_n_ent_percent_ps	average_number_of_named_entities_percent_per_s...
78	a_n_ent_money_ps	average_number_of_named_entities_money_per_sen...
79	a_n_ent_quantity_ps	average_number_of_named_entities_quantity_per_...
80	a_n_ent_ordinal_ps	average_number_of_named_entities_ordinal_per_s...
81	a_n_ent_cardinal_ps	average_number_of_named_entities_cardinal_per_...
82	simp_adj_var	simple_adjectives_variation
83	simp_adp_var	simple_adpositions_variation
84	simp_adv_var	simple_adverbs_variation
85	simp_aux_var	simple_auxiliaries_variation
86	simp_cconj_var	simple_coordinating_conjunctions_variation
87	simp_det_var	simple_determiners_variation
88	simp_intj_var	simple_interjections_variation
89	simp_noun_var	simple_nouns_variation
90	simp_num_var	simple_numerals_variation
91	simp_part_var	simple_particles_variation
92	simp_pron_var	simple_pronouns_variation
93	simp_propn_var	simple_proper_nouns_variation
94	simp_punct_var	simple_punctuations_variation
95	simp_sconj_var	simple_subordinating_conjunctions_variation
96	simp_sym_var	simple_symbols_variation
97	simp_verb_var	simple_verbs_variation
98	simp_space_var	simple_spaces_variation
99	root_adj_var	root_adjectives_variation
100	root_adp_var	root_adpositions_variation
101	root_adv_var	root_adverbs_variation
102	root_aux_var	root_auxiliaries_variation
103	root_cconj_var	root_coordinating_conjunctions_variation
104	root_det_var	root_determiners_variation
105	root_intj_var	root_interjections_variation
106	root_noun_var	root_nouns_variation
107	root_num_var	root_numerals_variation
108	root_part_var	root_particles_variation
109	root_pron_var	root_pronouns_variation

Table 19: Linguistic Features (0-109)

	Key	Feature-name
110	root_propn_var	root_proper_nouns_variation
111	root_punct_var	root_punctuations_variation
112	root_sconj_var	root_subordinating_conjunctions_variation
113	root_sym_var	root_symbols_variation
114	root_verb_var	root_verbs_variation
115	root_space_var	root_spaces_variation
116	corr_adj_var	corrected_adjectives_variation
117	corr_adp_var	corrected_adpositions_variation
118	corr_adv_var	corrected_adverbs_variation
119	corr_aux_var	corrected_auxiliaries_variation
120	corr_cconj_var	corrected_coordinating_conjunctions_variation
121	corr_det_var	corrected_determiners_variation
122	corr_intj_var	corrected_interjections_variation
123	corr_noun_var	corrected_nouns_variation
124	corr_num_var	corrected_numerals_variation
125	corr_part_var	corrected_particles_variation
126	corr_pron_var	corrected_pronouns_variation
127	corr_propn_var	corrected_proper_nouns_variation
128	corr_punct_var	corrected_punctuations_variation
129	corr_sconj_var	corrected_subordinating_conjunctions_variation
130	corr_sym_var	corrected_symbols_variation
131	corr_verb_var	corrected_verbs_variation
132	corr_space_var	corrected_spaces_variation
133	simp_ttr	simple_type_token_ratio
134	root_ttr	root_type_token_ratio
135	corr_ttr	corrected_type_token_ratio
136	bilog_ttr	bilogarithmic_type_token_ratio
137	uber_ttr	uber_type_token_ratio
138	simp_ttr_no_lem	simple_type_token_ratio_no_lemma
139	root_ttr_no_lem	root_type_token_ratio_no_lemma
140	corr_ttr_no_lem	corrected_type_token_ratio_no_lemma
141	bilog_ttr_no_lem	bilogarithmic_type_token_ratio_no_lemma
142	uber_ttr_no_lem	uber_type_token_ratio_no_lemma
143	n_adj	total_number_of_adjectives
144	n_adp	total_number_of_adpositions
145	n_adv	total_number_of_adverbs
146	n_aux	total_number_of_auxiliaries
147	n_cconj	total_number_of_coordinating_conjunctions
148	n_det	total_number_of_determiners
149	n_intj	total_number_of_interjections
150	n_noun	total_number_of_nouns
151	n_num	total_number_of_numerals
152	n_part	total_number_of_particles
153	n_pron	total_number_of_pronouns
154	n_propn	total_number_of_proper_nouns
155	n_punct	total_number_of_punctuations
156	n_sconj	total_number_of_subordinating_conjunctions
157	n_sym	total_number_of_symbols
158	n_verb	total_number_of_verbs
159	n_space	total_number_of_spaces
160	n_uadj	total_number_of_unique_adjectives
161	n_uadp	total_number_of_unique_adpositions
162	n_uadv	total_number_of_unique_adverbs
163	n_uaux	total_number_of_unique_auxiliaries
164	n_ucconj	total_number_of_unique_coordinating_conjunctions
165	n_udet	total_number_of_unique_determiners

	Key	Feature-name
166	n_uintj	total_number_of_unique_interjections
167	n_unoun	total_number_of_unique_nouns
168	n_unum	total_number_of_unique_numerals
169	n_upart	total_number_of_unique_particles
170	n_upron	total_number_of_unique_pronouns
171	n_upropn	total_number_of_unique_proper_nouns
172	n_upunct	total_number_of_unique_punctuations
173	n_usconj	total_number_of_unique_subordinating_conjunctions
174	n_usym	total_number_of_unique_symbols
175	n_uverb	total_number_of_unique_verbs
176	n_uspace	total_number_of_unique_spaces
177	a_adj_pw	average_number_of_adjectives_per_word
178	a_adp_pw	average_number_of_adpositions_per_word
179	a_adv_pw	average_number_of_adverbs_per_word
180	a_aux_pw	average_number_of_auxiliaries_per_word
181	a_cconj_pw	average_number_of_coordinating_conjunctions_pe...
182	a_det_pw	average_number_of_determiners_per_word
183	a_intj_pw	average_number_of_interjections_per_word
184	a_noun_pw	average_number_of_nouns_per_word
185	a_num_pw	average_number_of_numerals_per_word
186	a_part_pw	average_number_of_particles_per_word
187	a_pron_pw	average_number_of_pronouns_per_word
188	a_propn_pw	average_number_of_proper_nouns_per_word
189	a_punct_pw	average_number_of_punctuations_per_word
190	a_sconj_pw	average_number_of_subordinating_conjunctions_p...
191	a_sym_pw	average_number_of_symbols_per_word
192	a_verb_pw	average_number_of_verbs_per_word
193	a_space_pw	average_number_of_spaces_per_word
194	a_adj_ps	average_number_of_adjectives_per_sentence
195	a_adp_ps	average_number_of_adpositions_per_sentence
196	a_adv_ps	average_number_of_adverbs_per_sentence
197	a_aux_ps	average_number_of_auxiliaries_per_sentence
198	a_cconj_ps	average_number_of_coordinating_conjunctions_pe...
199	a_det_ps	average_number_of_determiners_per_sentence
200	a_intj_ps	average_number_of_interjections_per_sentence
201	a_noun_ps	average_number_of_nouns_per_sentence
202	a_num_ps	average_number_of_numerals_per_sentence
203	a_part_ps	average_number_of_particles_per_sentence
204	a_pron_ps	average_number_of_pronouns_per_sentence
205	a_propn_ps	average_number_of_proper_nouns_per_sentence
206	a_punct_ps	average_number_of_punctuations_per_sentence
207	a_sconj_ps	average_number_of_subordinating_conjunctions_p...
208	a_sym_ps	average_number_of_symbols_per_sentence
209	a_verb_ps	average_number_of_verbs_per_sentence
210	a_space_ps	average_number_of_spaces_per_sentence
211	fkre	flesch_kincaid_reading_ease
212	fkgl	flesch_kincaid_grade_level
213	fogi	gunning_fog_index
214	smog	smog_index
215	cole	coleman_liau_index
216	auto	automated_readability_index
217	rt_fast	reading_time_for_fast_readers
218	rt_average	reading_time_for_average_readers
219	rt_slow	reading_time_for_slow_readers

Table 20: Linguistic Features (110-219)

KALIMBA: Knowledge-Assisted Literature Mining for Biological Interaction Analysis

Niloofar Arazkhani[1], Maciej Kotecki[2], Brent Cochran[2], Natasa Miskov-Zivanov[1]

[1]University of Pittsburgh, Pittsburgh, PA, USA, [2] Tufts University, Medford, MA, USA

Correspondence: nmzivanov@pitt.edu

Abstract

The exponential growth of biomedical literature has made manual curation of biological interaction networks increasingly difficult. Existing automated biological interaction extraction systems address the scaling challenge but treat extraction as a final step, delivering structured output with limited or no integrated support for biologists to interactively verify, correct and contextually interrogate extracted interactions against their source evidence within the same environment. We present Knowledge-Assisted Literature Mining for Biological Interaction Analysis (KALIMBA), an end-to-end, human-in-the-loop platform that integrates three complementary extraction methods (NLP-only, LLM-only, and hybrid) alongside expert annotation and evidence-grounded conversational querying through retrieval-augmented generation (RAG) chat module driven by a dual-context prompt, within a single unified workflow. Evaluation on a corpus of 40 signaling-focused papers demonstrates that the LLM-only back-end recovers substantially more interactions than the NLP-only approach. RAG chat evaluation by a domain expert confirms that the conversational module provides scientifically grounded responses that support curation decisions beyond what the structured interaction table alone conveys.

1 Introduction

The curation of large-scale biological knowledge graphs is an important challenge in systems biology, as these structured representations of molecular interactions underpin mechanistic modeling of signaling pathways, disease mechanisms, and drug response Lo Surdo et al. (2017); Milacic et al. (2024); Oughtred et al. (2019); Türei et al. (2021). These knowledge graphs are populated from interactions reported across tens of thousands of primary research articles Szklarczyk et al. (2016). Databases such as SIGNOR Lo Surdo et al. (2017), PathwayCommons Rodchenkov et al. (2020), and BioGRID Oughtred et al. (2021) represent decades of expert curation, yet the rate of new discoveries consistently outpaces the capacity of manual curation to keep them current.

Significant research effort has been devoted to automating the extraction of biological interactions from the primary literature. Rule-based natural language processing (NLP) systems such as REACH Valenzuela-Escárcega et al. (2015), and INDRA Gyori et al. (2017) have demonstrated the feasibility of extracting structured regulatory interactions from biomedical articles at scale. The INDRA system in particular has established a strong foundation by integrating multiple reading engines and databases, and producing interactions in a standardized, machine-readable format with associated belief scores, which represent the confidence that a relationship is correct.

Despite this progress, existing systems share a common limitation: interaction extraction is their final step and there is no interface for annotating the quality or completeness of individual fields, or for asking follow-up questions about the context of reported findings. This disconnect between automated extraction and domain-expert judgment represents a significant barrier to further adoption of text mining tools in curation workflows.

Large language models (LLMs) offer a compelling opportunity to close the gap simultaneously from both ends. On the extraction side, LLMs provide broader coverage than rule-based NLP systems. On the curation side, the same conversational capabilities that make LLMs effective at extraction also make them natural assistants for post-extraction verification: a domain expert can query the source evidence behind a specific interaction, ask follow-up questions about experimental conditions, and receive grounded, context-aware responses without ever leaving the curation interface. A principled hybrid approach that combines the precision of established NLP systems with the

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 880–889
July 3-4, 2026 ©2026 Association for Computational Linguistics

coverage and conversational capabilities of LLMs, while keeping a domain expert in the loop throughout the entire workflow, has not yet been realized in a single integrated platform.

In this paper, we introduce Knowledge-Assisted Literature Mining for Biological Interaction Analysis (KALIMBA), a human-in-the-loop curation platform that addresses these limitations by unifying automated interaction extraction, expert annotation, and evidence-grounded conversational querying within a single integrated workflow. KALIMBA treats knowledge curation as a collaborative process between automated systems and domain experts, keeping the expert actively engaged throughout the entire workflow.

2 Related work

Biological interaction and relation extraction. Early approaches to automated extraction of biological interactions relied on rule-based pattern matching and dependency parse trees to identify regulatory relationships between named entities in biomedical text. The REACH Valenzuela-Escárcega et al. (2015) system demonstrated that rule-based event extraction could recover mechanistic interactions from biomedical abstracts and full-text articles. INDRA Gyori et al. (2017) built on this foundation by integrating multiple reading engines into a unified assembly pipeline that produces computable statements with associated belief scores, enabling downstream pathway modeling and network assembly.

The emergence of language models has substantially expanded the capabilities of biomedical information extraction. BioBERT Lee et al. (2020) is a domain-specific language model pre-trained on large-scale biomedical corpora that significantly outperforms general-domain BERT Devlin et al. (2019) on biomedical named entity recognition and relation extraction.

Human-in-the-loop curation tools. Human-in-the-loop approaches have been widely applied in biomedical knowledge curation, ranging from active learning frameworks that reduce annotation effort to web-based manual annotation interfaces designed for structured text labeling. BRAT Stenetorp et al. (2012) is an intuitive web-based tool for text annotation supported by NLP technology and has been widely adopted for constructing gold-standard evaluation datasets across a range of biomedical NLP tasks. More recently, OntoChat Zhang et al. (2024) demonstrated that conversational language model agents can meaningfully assist domain experts throughout a knowledge engineering workflow without requiring programming expertise. KALIMBA extends this principle to biological interaction curation, applying conversational LLM assistance to the verification and contextual exploration of automatically extracted interactions grounded in primary literature evidence.

3 Methods

3.1 System Overview

KALIMBA is an end-to-end, human-in-the-loop platform for the automated extraction, structured curation, and interactive verification of biological regulatory interactions from primary biomedical literature. The system comprises two functional layers: an automated literature retrieval and extraction layer supporting three complementary methods, and a human-in-the-loop curation layer integrating expert annotation, RAG-based conversational querying, carefully engineered prompt template, and data import and export. A schematic overview of the full system architecture is presented in Figure 1.

3.2 Literature Retrieval

KALIMBA supports two entry modes that determine how paper content enters the pipeline. In search mode, the user queries PubMed directly through the interface, selects papers from the returned results, and abstracts are cached as part of the search session before extraction begins. In import mode, the user loads a previously extracted BioRECIPE Holtzapple et al. (2024) interaction list directly into the curation layer, bypassing extraction entirely and proceeding immediately to annotation, quality rating, and RAG-based conversational querying.

BioRECIPE is a structured biological interaction representation in a human-readable format, and KALIMBA's import and export including its built-in annotation tool support BioRECIPE format. BioRECIPE has 28 fields, including regulator and regulated entity names, interaction sign, mechanism, cellular compartment, cell line, tissue type, and supporting evidence statement.

For search mode, paper content is retrieved from National Center for Biotechnology Information (NCBI) through a cascading three-step strategy using the Entrez Utilities (E-utilities) RESTful

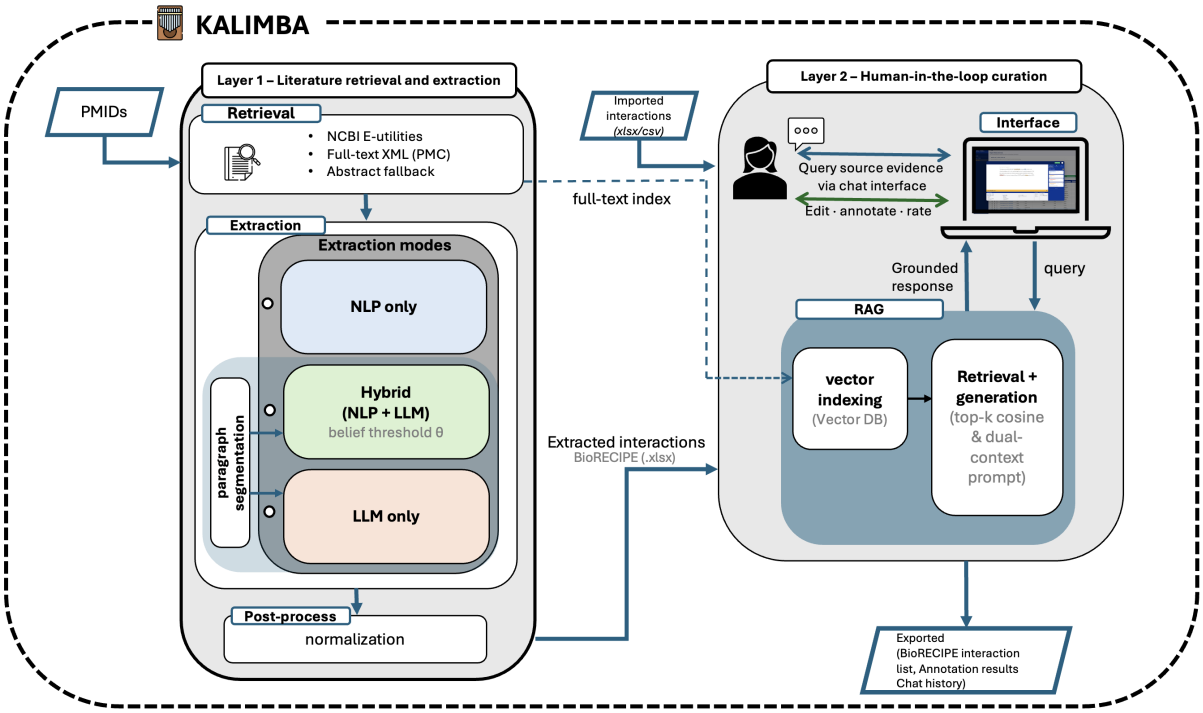


Figure 1: Overview of the KALIMBA framework

API Kans (2025). First, an elink call maps all input PMIDs simultaneously to their corresponding PubMed Central (PMC) identifiers; only PMIDs yielding a valid PMC entry are carried forward. Second, a single batch efetch call downloads full-text XML for all linked articles in one request. Body content is extracted by collecting all <p> elements within the <body> tag and joining them with double-newline separators to preserve paragraph boundaries. Third, a conditional efetch call retrieves PubMed abstracts for papers lacking PMC full-text coverage; this call is skipped entirely when abstracts are already cached from the search session, reducing total API calls to two for the common case.

Full-text content retrieved from PMC serves exclusively as the document corpus for RAG vector indexing (Section 3.7) and is not used as input to the extraction pipeline. Interaction extraction operates uniformly on abstract text for all input papers, regardless of full-text availability. This design choice ensures that extracted evidence statements are grounded in a consistent, universally available text source, preventing systematic differences in extraction quality between papers with and without PMC full-text coverage. In import mode, full-text content is fetched on demand at chat time if no index exists for the imported interaction's source papers, ensuring that RAG-based conversational querying remains available regardless of entry mode.

3.3 Paragraph Segmentation

Prior to extraction using LLM, each abstract is segmented into focused text units to constrain the language model's context window and to ensure that the Statements attribute of each extracted interaction is populated with a precise, traceable passage from the source text. This paragraph-level design serves two purposes: it focuses each extraction call on a narrow, semantically coherent passage, and it guarantees that the verbatim source text used by the annotator in the curation layer (Section 3.6) corresponds exactly to the context seen by the extraction model.

3.4 Interaction Extraction

KALIMBA exposes three configurable extraction methods, selectable at runtime via the interface (Appendix A).

NLP-only mode: Interactions are extracted via the INDRA reading pipeline, which uses NLP readers to process paper contents.

LLM-only mode: Each abstract paragraph produced by the segmentation stage (Section 3.3) is submitted independently to a language model with a structured prompt (Figure 2). KALIMBA is designed to be model-agnostic, supporting both open-

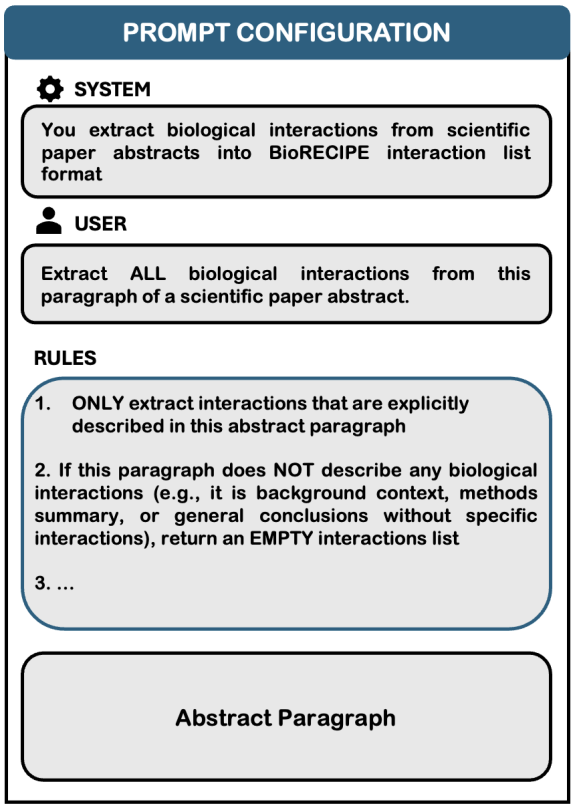


Figure 2: System and user prompt template for paragraph-level extraction.

source models served through the Ollama API and any OpenAI-compatible remote endpoint, allowing researchers to select the language model that best fits their computational resources, data privacy requirements, and performance needs.

Hybrid mode: The rule-based NLP-driven tool is applied first across all input papers. Interactions meeting a configurable belief score threshold (default $\theta = 0.7$) are retained directly. Papers with no NLP output, and papers with interactions that fall entirely below the belief threshold, are passed to the LLM extraction path. If LLM extraction subsequently fails for any paper in this set, the low-confidence interactions are preserved as a fallback rather than discarded. This strategy maximizes recall by using the LLM's broad coverage while preserving the precision of high-confidence traditional NLP extractions, avoiding the need to choose between the two approaches. The balance between NLP and LLM contributions is controlled by the belief threshold θ, which is developed as a continuous slider in the interface (range 0-1) that biologists can adjust before initiating extraction. This design gives biologists direct, interpretable control over the precision-recall trade-off without requiring any knowledge of the underlying implementation.

3.5 Post-Processing

All extracted interactions undergo post-processing normalization, which resolves non-standard field values and recovers a subset of missing fields through controlled vocabulary mapping and external database lookup. Type attributes are mapped to a controlled vocabulary of six categories: protein, gene, chemical, RNA, protein family, and biological process. Biological database identifiers are validated against type-appropriate sources, UniProt UniProt Consortium (2015) for proteins, HGNC Seal et al. (2023) for genes, PubChem Kim et al. (2016) for chemicals, Ensembl Harrison et al. (2024) for RNA, Pfam Mistry et al. (2021) for protein families, and Gene Ontology Gene Ontology Consortium (2019) for biological processes. Subcellular compartments are resolved and standardized against a set of seven canonical cellular locations, each matched with its corresponding Gene Ontology component ID. Mechanisms are similarly standardized by matching extracted values against a controlled vocabulary of 13 terms. Entity names unresolved during extraction are looked up asynchronously against the UniProt REST API (proteins and genes), HGNC (gene symbols), PubChem (chemicals), and the EBI QuickGO service (biological processes). Rows lacking either a regulator or regulated entity name, or missing an evidence statement, are discarded. The Source attribute is set programmatically to reflect the extraction origin of each row.

3.6 Human-in-the-Loop Curation and Annotation

A distinguishing feature of KALIMBA is its integrated human-in-the-loop curation layer, which treats automated extraction not as a final product but as a first-pass draft, subject to expert review and correction. The curation interface is implemented as an in-browser interaction table that uses the BioRECIPE interaction representation format with inline editing, per-row deletion, and column-level text filtering.

Each extracted interaction is linked to the verbatim paragraph text submitted to the extraction model, providing a traceable evidence statement against which automated field values can be verified. Clicking any statement opens the text annotation interface, where the evidence sentence is displayed with span highlights already pre-populated

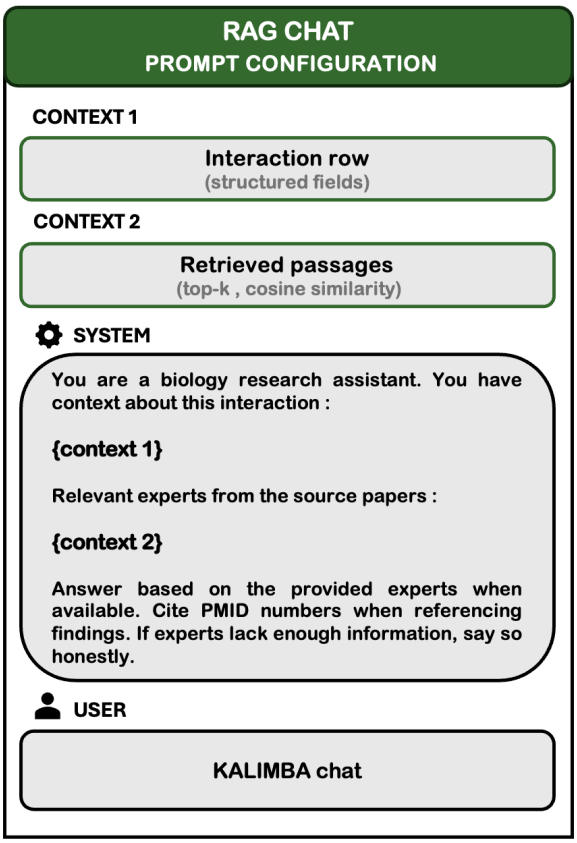


Figure 3: Dual-context prompt structure for RAG-based conversational querying.

from the extracted field values. From there, the biologist can verify whether each highlighted span correctly corresponds to its assigned BioRECIPE attribute, adjust any incorrect highlights, or add new ones for fields that the model left empty.

The interface also supports per-field quality rating and commentary, an overall interaction-level quality score, and a free-text general note field. The resulting structured evaluation data serves as a rich human feedback signal for downstream model evaluation and fine-tuning.

3.7 Retrieval-Augmented Interaction Chat

While interaction databases usually capture rich contextual information such as cell type, tissue, or experimental condition, domain experts often need to return to the primary literature to interrogate the full experimental context behind a specific interaction. KALIMBA addresses this through a tightly integrated RAG chat module that enables domain experts to query the source literature in natural language, anchored to a specific extracted interaction.

During the extraction phase in chat module, full-text content retrieved from PMC is segmented into overlapping 500-word chunks with a 100-word stride, preserving local semantic coherence while enabling retrieval at sub-section granularity. Each chunk is encoded into a dense vector representation using a sentence embedding model and stored in a vector database indexed by cosine similarity.

For papers unavailable in PMC, abstract text is indexed as a fallback. If no pre-built index exists at query time, for instance, when a user initiates chat on an interaction imported from an external file rather than extracted within the current session, full-text fetching and indexing are triggered on demand.

At query time, the user's natural language message is encoded using the same model and the top-k most similar passages (default k = 5) are retrieved by approximate nearest-neighbor search over the session collection. Retrieved passages are injected into the language model's system prompt alongside the structured format with attributes of the selected interaction (Figure 3). This dual-context design grounds responses simultaneously in the structured semantics of the BioRECIPE format (context 1) and in the unstructured experimental narrative of the source paper (context 2), reducing the risk of hallucinated mechanistic claims that are not supported by the cited evidence.

This capability represents a qualitatively different mode of engagement with curated biological knowledge. A researcher inspecting an extracted phosphorylation interaction can immediately ask: "What experimental method was used to validate this interaction?" or " Did the authors note any dose-dependent effects?". These questions can directly help curator in completing and enriching structured interaction records.

3.8 Data Import and Export

To ensure compatibility with a range of systems biology tools Kanehisa et al. (2016); Luo et al. (2026); Milacic et al. (2024); Pratt et al. (2015) and to support KALIMBA as a standalone curation environment, the system includes an import layer accepting BioRECIPE formatted files.

KALIMBA provides three structured export formats. The interaction list export produces a single-sheet Excel workbook with one column per BioRECIPE attribute, suitable for direct use in downstream pathway modeling tools. The annotation export produces a four-sheet workbook capturing span-level annotation details, per-interaction summaries, per-attribute quality ratings, and a full rating matrix. The chat history export produces a

two-sheet workbook containing message-level dialogues and conversation summaries, with each row enriched with the full BioRECIPE context of the discussed interaction.

Methods	Total interactions	Papers	
		covered	no output
NLP-only (INDRA)	392	24	16
LLM-only (LLaMA 3.1)	363	39	1
Hybrid ($\theta = 0.7$)	388	39	1

Table 1: Extraction coverage across methods.

Source	Extraction origin	Interactions	
		count	percentage
LLM	papers INDRA missed	191	49.2%
	low-confidence INDRA, score < 0.7	161	41.5%
INDRA /REACH	high-confidence, score ≥ 0.7, retained	32	8.2%
INDRA /Sparser	high-confidence, score ≥ 0.7, retained	4	1.0%
Total		**388**	**100%**

Table 2: Distribution of hybrid mode interactions.

4 Experiments

4.1 Experimental Setup

Dataset. To evaluate KALIMBA's extraction capabilities, we used a corpus of 40 signaling-focused biomedical papers from high-impact journals including Cell and Nature, published between 2023 and 2026. selected by a domain expert. The corpus is biased toward the RAS and JAK signaling pathways while maintaining diversity across related signaling contexts. Papers involving immune cell signaling were excluded due to the subtleties of cell-type-specific signaling that introduce additional ambiguity beyond the scope of this evaluation.

Models. For the LLM-only and hybrid extraction, we used a locally hosted open-source model (LLaMA 3.1 8B served via the Ollama API) Grattafiori et al. (2024). Although KALIMBA supports remote OpenAI-compatible endpoints such as GPT-4o Hurst et al. (2024), we used LLaMA configuration in this evaluation to avoid API costs and to demonstrate that KALIMBA produces meaningful extraction results without requiring access to closed-source API-dependent models. All experiments were conducted on a MacBook Pro with an Apple M3 Pro chip and 18 GB memory.

Metrics. The 40 papers were processed through all three extraction methods independently. Extraction outputs were compared across methods in terms of paper coverage, total interaction yield, inter-mode agreement, and confidence score distribution. For the RAG chat evaluation, a domain expert curated a set of standardized questions per interaction and rated the system responses on a 0-to-5-star scale, providing qualitative notes for each response.

4.2 Extraction Performance

Table 1 reports the total number of interactions extracted per mode and the number of papers for which at least one interaction was recovered across the full 40-paper corpus. The NLP-only back-end failed to produce any interactions for 16 of the 40 papers. All 16 of these papers were successfully processed by the LLM-only method, demonstrating that the LLM provides substantial coverage recovery for papers outside NLP's effective range. The hybrid mode inherits the LLM's coverage advantage, recovering interactions from 39 of 40 papers while retaining high-confidence NLP extractions where available.

Table 2 reports the source breakdown of interactions in the hybrid output, characterizing the practical contribution of each component.

Only 9.2% of hybrid interactions were contributed directly by high-confidence NLP output. The remaining 90.7% were produced by LLM re-extraction either for papers NLP missed entirely (49.2%) or for papers where NLP's belief scores fell below $\theta = 0.7$ (41.5%). This distribution confirms that the hybrid routing strategy is functioning as designed: NLP's precision is preserved for the interactions it handles confidently, while the LLM fills in coverage for the substantial portion of the corpus that NLP either cannot process or does not assign high confidence scores.

4.3 Confidence and Inter-Mode Agreement

Table 3 reports the confidence score distribution per method. NLP-only scores cluster narrowly between 0.650 and 0.980, reflecting NLP's con-

Method	Mean	Median	Min	Max
NLP-only (INDRA)	0.67	0.65	0.65	0.98
LLM-only (LLaMA 3.1)	0.87	1.00	—	1.00
Hybrid ($\theta = 0.7$)	0.85	0.90	0.00	1.00

Table 3: Confidence score distribution per method.

servative belief scoring system. The LLM-only back-end assigns higher scores on average, but these scores are self-reported and do not guarantee extraction correctness. A limitation directly observed in the RAG chat evaluation, where a highly confident but incorrectly extracted interaction received zero stars across all expert questions.The hybrid mode achieves a higher mean score than NLP-only alone as a direct consequence of the low-confidence NLP interactions are replaced by LLM re-extractions that receive higher confidence scores. Table 4 reports pairwise interaction overlap across the three methods, which show remarkably low agreement, with only 9 interactions agreed upon by all three methods. These 9 interactions include canonical RAS pathway relationships such as MEK→ERK and RASA1→RAS, providing a validity check that all three methods correctly capture well-established signaling interactions. The substantially higher LLM–Hybrid overlap (129 pairs) compared to NLP–LLM overlap (10 pairs) reflects the fact that 90.7% of hybrid interactions are LLM-derived, making these two the most similar pair. The low overall consensus confirms that the three methods are largely complementary rather than redundant, each recovering a distinct subset of the interactions present in the corpus and strengthening the case for the hybrid design.

4.4 RAG Chat Quality Evaluation

To evaluate the RAG-based conversational querying module, a domain expert curated a set of six standardized questions that were defined targeting distinct aspects of the supporting evidence.

Each response was rated by the domain expert on a 0-to-5-star scale and followed by a qualitative curator note where responses were incomplete or incorrect. A total of 47 responses were evaluated across 8 interactions. Table 5 shows that the overall mean rating of 2.64 out of 5 is largely driven down by one incorrectly extracted interaction that received 0.00 across all questions. Excluding it, the

Comparison	Shared pairs	Unique to first	Unique to second
NLP-only vs LLM-only	10	325	231
NLP-only vs Hybrid	44	325	221
LLM-only vs Hybrid	129	231	221
Agreed by all three	9	—	—

Table 4: Pairwise interaction overlap across methods (regulator and regulated name matches, n=40 papers).

adjusted mean rises to 3.27 out of 5 with 83.7% of responses rated 3 stars or above, indicating strong performance for correctly extracted interactions

Qualitative findings. Expert review of curator notes revealed three recurring patterns. First, RAG chat quality is tightly coupled to extraction correctness. The interaction AMP-activated protein kinase → glucose metabolism received a rating of 0.00 across all six questions because the interaction direction was incorrectly extracted. This finding suggests that the chat module itself can serve as a verification tool when responses are inconsistent with biological expectations. This further signals to the expert that the underlying extracted interaction may be incorrect and warrants correction in the annotation layer before further querying.

Second, the system performs strongest on directional and pathway-level questions. Activation vs. inhibition questions received the highest mean rating (3.50) and signaling pathway context questions the second highest (3.12). These question types are most directly answerable from the dual-context prompt that combines the structured BioRECIPE row with the retrieved passages, and the results suggest this design provides reliable grounding for sign and pathway membership queries.

Third, mechanistic and experimental evidence questions are more challenging. Biological mechanism (2.12) and cell line (2.00) questions received the lowest ratings. Expert notes identified two specific failure modes: responses occasionally cited literature referenced within retrieved passages rather than the experimental findings of the paper itself, and responses sometimes provided accurate but incomplete summaries of experimental evidence, omitting key details such as specific phosphorylation site identification, mutational analyses, and

tissue-specific knockout results.

Question type	Rating	
	mean	% rated ≥ 3 stars
Activation vs inhibition	3.50	75.0%
Signaling pathway context	3.12	87.5%
Explicit statement verification	2.75	62.5%
Experimental evidence	2.25	75.0%
Biological mechanism	2.12	62.5%
Cell line / condition	2.00	71.4%
Overall	2.64	72.3%

Table 5: Mean expert ratings of RAG chat responses by question type (n=47, scale 0–5).

Expert Utility Assessment. Following the quantitative rating evaluation, the domain expert was asked three follow-up questions regarding the utility of the RAG chat module for biological curation. The expert confirmed that the chat responses provided meaningful additional context beyond the structured interaction row across all evaluated interactions. Three specific advantages were identified. First, the chat module demonstrated the ability to correct upstream extraction errors for example for the incorrectly extracted interaction AMP-activated protein kinase → glucose metabolism, the expert noted that KALIMBA's responses consistently described the correct interaction direction, effectively flagging the extraction error without explicit prompting. Second, for interactions where the connection type was incorrectly assigned as direct or indirect, the chat module provided evidence that allowed the expert to identify and correct the misclassification. Third, in the cases where the structured interaction row did not include any information on experimental evidence, mechanism, site, cell line, cell type, or tissue type, the chat module provided varying degrees of detail across all of these attributes for every interaction, which the expert described as a significant advantage for curation.

The expert confirmed they would use the KALIMBA chat feature in their own literature review and curation workflow.

5 Conclusion

We presented KALIMBA, a human-in-the-loop curation platform that addresses a persistent gap in biomedical text mining: the disconnect between automated interaction extraction and the expert biological judgment required to verify, correct, and contextually interpret extracted knowledge. KALIMBA integrates three complementary extraction methods, a span-level annotation interface, and a retrieval-augmented conversational querying module within a single unified workflow. Together these capabilities treat biological knowledge curation as a collaborative process between automated systems and domain experts rather than a pipeline that terminates at extraction.

Evaluation on a 40-paper signaling-focused corpus revealed several findings with implications beyond KALIMBA itself. The NLP-only method failed to extract any interactions from 40% of the corpus, a coverage gap that the LLM method recovered entirely, demonstrating that LLM-based extraction provides meaningful complementarity to established NLP systems for recent literature. The hybrid mode successfully combined both approaches, with 90.7% of its output contributed by LLM re-extraction and only 9.3% retained directly from high-confidence NLP output, reflecting the corpus composition bias toward recently published papers underrepresented in NLP's knowledge base. Inter-mode agreement was remarkably low, only 9 interactions were agreed upon by all three methods confirming that the three methods are largely complementary rather than redundant, and that each recovers a distinct subset of the interactions present in the literature. The 9 consensus interactions included canonical RAS pathway relationships such as MEK→ERK and RASA1→RAS, providing validity that all three methods correctly capture well-established signaling interactions.

The RAG chat evaluation demonstrated that conversational querying of source evidence provides grounded responses for correctly extracted interactions, with an adjusted mean rating of 3.27 out of 5 and 83.7% of responses rated 3 stars or above when the one incorrectly extracted interaction is excluded. The finding that chat quality is tightly coupled to extraction quality reinforces the importance of the annotation and correction layer as a prerequisite for productive conversational querying. Together these results position KALIMBA as a practical tool for domain experts who need to extract, verify, and explore biological interactions from the primary literature at scale, without requiring computational expertise or access to commercial API services.

6 Limitations and Future Work

The current evaluation has several limitations that are considered for future work. The corpus of the papers that we used has 40 signaling-focused papers recommended by domain experts, which is very rich in terms of studying the signaling pathways, particularly in the RAS and JACK signals, but it is not constructed as a dataset of gold standard interaction extraction in BioRECIPE format, so the extraction results are reported in terms of coverage not in accuracy related metrics such as precision, recall and F1 against a gold standard dataset. Evaluating the accuracy of extractions using an expanded, more diverse corpus is a priority for our future work. For the RAG chat evaluation, we relied on a single-domain expert who rated 47 responses across 8 interactions using a structured set of question types. In future work, we will add more independent evaluators to strengthen the reliability of these findings and compute an inter-annotator score to generalize the results. To make the system more user-friendly, a formal usability evaluation of the interface with domain expert biologists is planned as future work.

The current prompt template follows a general instruction-based design. Future work will improve extraction performance by incorporating curated few-shot examples into the prompt and evaluating results using commercial models such as GPT-4o.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Gene Ontology Consortium. 2019. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. 2017. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology*, 13(11):MSB177651.

Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, and 1 others. 2024. Ensembl 2024. *Nucleic acids research*, 52(D1):D891–D899.

Emilee Holtzapple, Gaoxiang Zhou, Haomiao Luo, Difei Tang, Niloofar Arazkhani, Casey Hansen, Cheryl A Telmer, and Natasa Miskov-Zivanov. 2024. The biorecipe knowledge representation format. *ACS Synthetic Biology*, 13(8):2621–2624.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462.

Jonathan Kans. 2025. Entrez® direct: E-utilities on the unix command line. In *Entrez® Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, and 1 others. 2016. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Prisca Lo Surdo, Alberto Calderone, Gianni Cesareni, and Livia Perfetto. 2017. Signor: a database of causal relationships between biological entities—a short guide to searching and browsing. *Current protocols in bioinformatics*, 58(1):8–23.

Haomiao Luo, Casey Hansen, Niloofar Arazkhani, Cheryl A Telmer, Difei Tang, Gaoxiang Zhou, Peter Spirtes, and Natasa Miskov-Zivanov. 2026. Violin: A modular framework for scalable reconciliation of heterogeneous interaction graphs. *bioRxiv*, pages 2024–07.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, and 1 others. 2024. The reactome pathway knowledgebase 2024. *Nucleic acids research*, 52(D1):D672–D678.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, and 1 others. 2021. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419.

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, and 1 others. 2021. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, and 1 others. 2019. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541.

Dexter Pratt, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, Carol Miello, Lyndon Hicks, Sandor Szalma, and 1 others. 2015. Ndex, the network data exchange. *Cell systems*, 1(4):302–305.

Igor Rodchenkov, Ozgun Babur, Augustin Luna, Bulent Arman Aksoy, Jeffrey V Wong, Dylan Fong, Max Franz, Metin Can Siper, Manfred Cheung, Michael Wrana, and 1 others. 2020. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic acids research*, 48(D1):D489–D497.

Ruth L Seal, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A Bruford. 2023. Genenames. org: the hgnc resources in 2023. *Nucleic acids research*, 51(D1):D1003–D1009.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, and 1 others. 2016. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, and 1 others. 2021. Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular systems biology*, 17(3):MSB20209923.

UniProt Consortium. 2015. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212.

Marco A Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu, and Thomas Hicks. 2015. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, pages 127–132.

Bohui Zhang, Valentina Anita Carriero, Katrin Schreiberhuber, Stefani Tsaneva, Lucía Sánchez González, Jongmo Kim, and Jacopo de Berardinis. 2024. Ontochat: a framework for conversational ontology engineering using language models. In *European semantic web conference*, pages 102–121. Springer.

A Interface

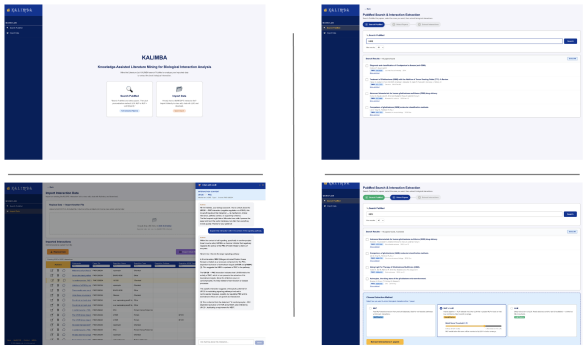


Figure 4: KALIMBA Interface – https://www.boheme.pitt.edu/kalimba

When Retrieval Doesn't Help: A Large-Scale Study of Biomedical RAG

Erfan Nourbakhsh, Rocky Slavin, Ke Yang, and Anthony Rios
The University of Texas at San Antonio
{Erfan.Nourbakhsh, Rocky.Slavin, Ke.Yang, Anthony.Rios}@utsa.edu

Abstract

Medical question answering is a high-stakes setting where factual errors can have serious consequences. Retrieval-augmented generation (RAG) is widely viewed as a promising solution, and prior work has reported substantial gains for large medical QA models. We revisit this assumption across a broad range of open-weight instruction-tuned models spanning 7B to 72B parameters. Across five models, ten biomedical QA datasets, four retrieval methods, and four retrieval corpora, we find that retrieval yields only small and inconsistent improvements over a no-retrieval baseline, typically within 1–2 points. In contrast, the choice of backbone model has a much larger effect than the choice of retriever or corpus, and expert and layman retrieval sources perform similarly in most settings. These results suggest that the main bottleneck is not retrieval quality alone, but the model's limited ability to use retrieved evidence effectively. Code is available here: https://github.com/erfan-nourbakhsh/BioMedicalRAG

1 Introduction

Accurate and reliable medical question answering is a high-stakes problem, where errors can have direct consequences for patient safety. Large language models (LLMs) have recently shown strong performance on a range of biomedical question answering tasks (Singhal et al., 2023; Hendrycks et al., 2021; Jin et al., 2021). However, they remain prone to hallucination, producing fluent but factually incorrect responses (Ji et al., 2023), and to knowledge staleness due to their reliance on fixed training corpora. In the medical domain, these limitations are especially problematic because even small factual errors can lead to harmful downstream decisions.

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has become a leading approach for addressing these limitations by grounding model outputs in retrieved external evidence. By incorporating supporting documents at inference time, RAG offers a mechanism for improving factuality, transparency, and access to more current knowledge. As a result, RAG has been adopted widely in biomedical NLP, where recent work has reported substantial gains from retrieval-based methods. For example, Xiong et al. (2024) showed that MedRAG improves biomedical QA accuracy by as much as 18% over chain-of-thought prompting, while Tang et al. (2024) found that multi-agent LLM systems can further improve medical reasoning performance. These findings have led to growing interest in retrieval-centered biomedical QA systems, with increasing attention to the choice of corpora, retrieval methods, and model backbones.

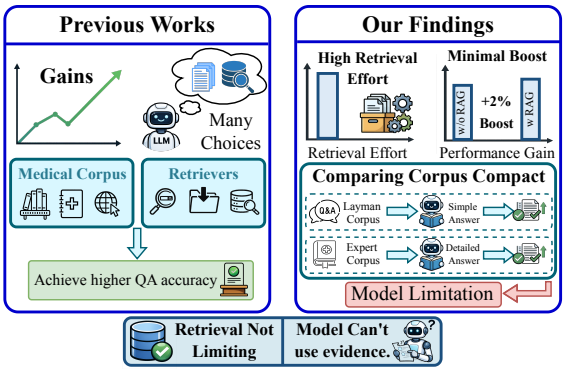


Figure 1: Overview of our motivation and main finding: across models from 7B to 72B, retrieval yields only small gains, suggesting that the main bottleneck is evidence use rather than retrieval quality.

However, an important gap remains. Prior systematic studies of medical RAG (Xiong et al., 2024) have largely focused on large proprietary or 70B-scale models (GPT-3.5, GPT-4, Mixtral-8x7B, Llama2-70B) under zero-shot multiple-choice evaluation, leaving unclear whether their gains carry over to 7B–8B models that are far more practical under real hardware constraints. Existing evaluations have also focused primarily on expert-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 890–910
July 3-4, 2026 ©2026 Association for Computational Linguistics

level biomedical questions, with little attention to consumer-health queries or community-generated retrieval sources.

In this paper, we revisit biomedical RAG under a substantially different and more comprehensive setting. We evaluate five open-weight instruction-tuned models spanning 7B to 72B parameters: Qwen2.5-7B-Instruct (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), LLaMA-3.1-70B-Instruct (Grattafiori et al., 2024), and Qwen2.5-72B-Instruct (Yang et al., 2025), across ten biomedical QA datasets spanning both lay and expert questions and covering both open-ended and multiple-choice formats. We compare four retrieval methods, BM25, TF-IDF, MedCPT, and Hybrid RRF, across four retrieval corpora, including both expert biomedical resources and consumer-facing health sources: PubMed abstracts, medical textbooks, Yahoo Answers, and HealthCareMagic. We also evaluate against a no-retrieval baseline in order to isolate the contribution of retrieval itself.

Our results challenge the prevailing picture from prior studies. Across all five models, retrieval yields only small and inconsistent gains: the gap between the best retrieval configuration and the no-retrieval baseline is usually within 1–2 points (e.g., BERTScore 62.88 vs. 61.72 for Llama-8B, 63.23 vs. 61.28 for Qwen-7B), and differences across retrieval corpora are similarly modest even for the larger 70B models. By contrast, backbone model choice has a much larger effect than retriever or corpus selection, and expert versus lay retrieval sources differ by less than 2 points in most settings. Figure 1 illustrates the key implication: the limiting factor is not retrieval quality but the generator's capacity to incorporate retrieved evidence.

Our contributions are: (1) A large-scale evaluation of biomedical RAG covering 5 models from 7B to 72B parameters, 10 QA datasets, 4 retrieval methods, and 4 retrieval corpora. (2) We show that retrieval yields only small and inconsistent improvements across all model scales (typically within 2 points), challenging the gains reported in prior large-model studies. (3) We show that backbone model choice matters more than retriever or corpus choice, and provide evidence that the main bottleneck is the model's weak use of retrieved evidence.

2 Related Work

Retrieval-Augmented Generation. RAG was introduced by Lewis et al. (2020) as a method to enhance language models on knowledge-intensive tasks by conditioning generation on documents retrieved from a non-parametric memory. The approach combines a parametric sequence-to-sequence model with a dense passage retrieval component (Karpukhin et al., 2020) and has been extended in numerous directions, including iterative retrieval (Trivedi et al., 2023; Shao et al., 2023), self-reflective retrieval (Asai et al., 2024), and query rewriting (Ma et al., 2023); for a broad survey of RAG paradigms and architectures, see Gao et al. (2023). In the biomedical domain, RAG has been applied to clinical decision support (Xiong et al., 2024), scientific literature search (Jin et al., 2023), and consumer health QA (Li et al., 2023). However, most prior biomedical RAG studies either lack systematic comparison across retrieval configurations or are limited in dataset coverage.

Benchmarking Medical RAG. The most directly related work to ours is the MIRAGE benchmark and MedRAG toolkit by Xiong et al. (2024), which evaluates 41 combinations of corpora, retrievers, and backbone LLMs on five medical QA datasets restricted to multiple-choice questions. MIRAGE shows that RAG can improve LLM accuracy by up to 18% and identifies PubMed combined with BM25 or MedCPT as strong retrieval configurations. However, MIRAGE exclusively uses zero-shot prompting and evaluates primarily large models (GPT-4, GPT-3.5, Mixtral-8×7B, Llama2-70B), leaving open the question of whether these gains hold for smaller, more widely deployable models. Concurrent large-model evaluations, such as Nori et al. (2023), who find that GPT-4 surpasses the USMLE passing threshold by over 20 points even without retrieval augmentation, further underscore that model scale is a critical confound in existing medical benchmarks. Tang et al. (2024) propose a zero-shot multi-agent framework achieving competitive GPT-4 performance on MMLU Medical, yet neither this nor MIRAGE examines retrieval for open-ended or consumer-health queries at the 7–8B scale. Shi et al. (2023) show that irrelevant retrieved passages can mislead LLMs, a concern especially acute for smaller models, while Ovadia et al. (2024) find retrieval augmentation outperforms knowledge fine-tuning primarily for large models, further motivating our cross-scale evalua-

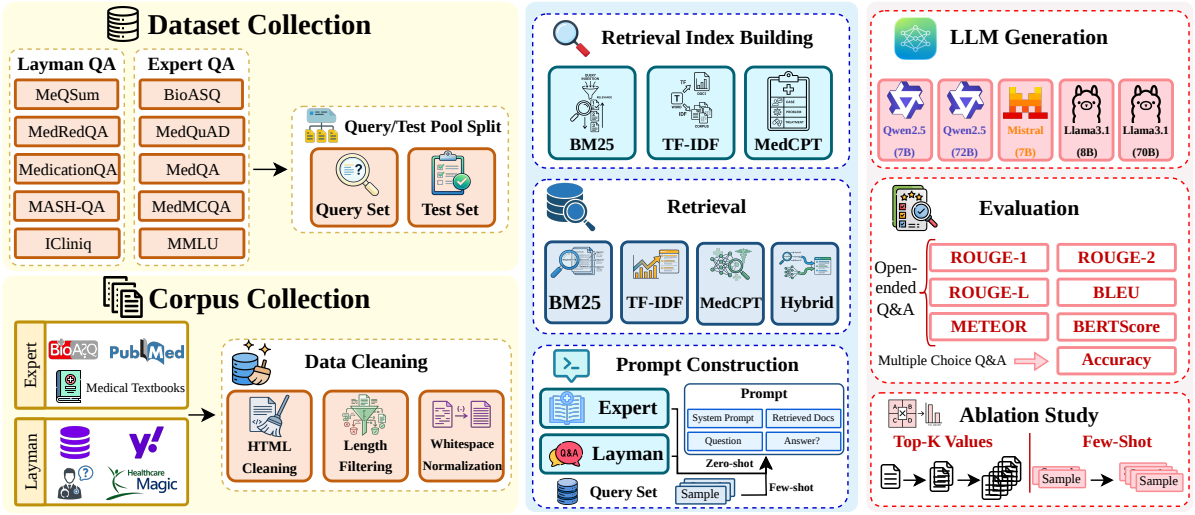


Figure 2: Experimental pipeline overview.

tion.

Biomedical Question Answering Datasets. Biomedical QA has long served as a testbed for evaluating NLP systems in medicine (Krithara et al., 2023; Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021). Expert-oriented benchmarks such as BioASQ (Nentidis et al., 2025), MedQA-USMLE (Jin et al., 2021), and MedMCQA (Pal et al., 2022) test clinical and examination-level knowledge, while consumer-health datasets such as MeQSum (Ben Abacha and Demner-Fushman, 2019a), MedRedQA (Nguyen et al., 2023), MedicationQA (Abacha et al., 2019), MASH-QA (Zhu et al., 2020), and ChatDoctor-iCliniq (Li et al., 2023) reflect more informal, everyday health information needs. MedQuAD (Ben Abacha and Demner-Fushman, 2019b) and MMLU Medical (Hendrycks et al., 2021) bridge the two groups by covering structured NIH-sourced QA and standardised medical knowledge. Despite this rich landscape, most RAG studies focus on MCQ-format expert benchmarks and omit the open-ended and layman query types that constitute the bulk of real-world health information needs, a gap we directly address.

Retrieval Methods in Biomedical NLP. Sparse retrieval methods have been dominant in biomedical information retrieval. BM25 (Robertson and Zaragoza, 2009), a probabilistic bag-of-words ranking function, remains a strong baseline and is adopted as the primary retriever in MedRAG (Xiong et al., 2024). TF-IDF (SPARCK JONES, 1972), a simpler precursor that models term specificity without BM25's saturation and length normalisation, provides a useful lower bound for sparse retrieval.

Dense retrieval with domain-adapted encoders has gained traction. MedCPT (Jin et al., 2023) was trained contrastively on large-scale PubMed search logs and demonstrates strong zero-shot biomedical retrieval, outperforming general-domain encoders on medical tasks. Fusion methods such as Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) combine sparse and dense ranked lists and have been shown to outperform individual retrievers without requiring additional training. While MedRAG includes RRF as a configuration, it does not systematically isolate the contribution of each component retriever across diverse query types and corpora, which we do in this study.

3 Experiments

Our experiments systematically compare sparse, dense, and hybrid retrieval strategies across four corpora, ten QA datasets spanning expert and layman health queries, and five open-weight instruction-tuned models ranging from 7B to 72B parameters. Figure 2 provides a visual overview of the full experimental pipeline.

Evaluation Dataset and Knowledge Base.

Evaluation Datasets. We evaluate across ten biomedical and consumer-health question answering datasets grouped into two user types: *layman* datasets reflecting everyday consumer-health language, and *expert* datasets targeting biomedical professionals or medical students. Dataset statistics and split sizes are summarised in Table 7 in Appendix A. For all datasets, examples lacking a

question or a reference answer are discarded before any split is finalised, and whenever random sampling is needed it is performed with a fixed seed of 42.

Layman datasets. **MeQSum** (Ben Abacha and Demner-Fushman, 2019a) contains 1,000 consumer health questions from the U.S. National Library of Medicine. Following Zhang et al. (2022), we reserve 500 examples for evaluation and use the remaining 500 as the few-shot query pool. **MedRedQA** (Nguyen et al., 2023) provides over 51,000 consumer question–physician answer pairs from Reddit's `/r/AskDocs`; we sample 1,000 evaluation examples from the official test split (5,099 examples) and combine the training (40,792) and validation (5,100) splits into the query pool. **MedicationQA** (Abacha et al., 2019) contains 690 real consumer medication questions; we randomly sample 500 for evaluation and retain the remaining 189 as the query pool. **MASH-QA** (Zhu et al., 2020) offers over 34,000 WebMD-derived healthcare Q&A pairs; we randomly sample 1,000 examples from the official test file (2,614 entries) and use the full training set (19,989 examples) as the query pool. **ChatDoctor-iCliniq** (Li et al., 2023) comprises 7,321 real patient–physician conversations from iCliniq.com; we randomly sample 1,000 for evaluation and retain the remaining 6,321 as the query pool.

Expert datasets. **BioASQ Task B** (Nentidis et al., 2025) provides expert biomedical questions grounded in PubMed literature; following the official benchmark protocol, we use the Task 13B golden test set (restricted to summary-type questions, 80 examples) and the Task 13B training set (1,283 examples) as the query pool. **MedQuAD** (Ben Abacha and Demner-Fushman, 2019b) contains 47,457 medical Q&A pairs from 12 NIH websites, of which 16,407 are publicly available; we randomly sample 1,000 for evaluation and use the remaining 15,407 as the query pool. **MedQA-USMLE** (Jin et al., 2021) provides USMLE clinical vignette MCQs; we use the official test split (1,273 examples) for evaluation and the official training split (10,178 examples) as the query pool. **MedMCQA** (Pal et al., 2022) contains 194k+ MCQs from AIIMS and NEET PG medical entrance exams; as the official test split is unlabelled, we randomly sample 1,000 from the validation set (6,150 examples) for evaluation and use the full training set (182,822 examples) as the query pool. **MMLU Medical** (Hendrycks et al., 2021): following Tang et al. (2024), we restrict to six medical sub-tasks, *anatomy*, *clinical_knowledge*, *college_biology*, *college_medicine*, *medical_genetics*, and *professional_medicine*, totalling 1,242 examples. Roughly 100 examples per sub-task (600 in total) are used for evaluation; the remaining 642 form the query pool.

Knowledge Bases. We build four retrieval corpora covering both expert biomedical and layman health domains, as summarised in Table 8 in Appendix A. All corpora are indexed as whole records without further chunking. For Q&A-style corpora (Yahoo Answers and HealthCareMagic), each document concatenates the question or title with the corresponding answer body.

BioASQ / PubMed (Krithara et al., 2023) consists of 16.2 million PubMed abstracts with human-assigned MeSH annotations and serves as the primary expert-domain knowledge base.

Medical Textbooks (Xiong et al., 2024) provides 125,847 retrieval-friendly chunks ($\leq$1,000 characters each) drawn from 18 authoritative biomedical textbooks spanning anatomy, physiology, pharmacology, pathology, and clinical medicine.

Yahoo Answers (Yahoo! Research, 2009) is an open-domain community Q&A corpus; from the original 1.4 million records we retain 1,238,506 after quality filtering, discarding entries whose answer body contains fewer than five words or whose combined question–answer text falls below ten words.

HealthCareMagic (Li et al., 2023) contains 112,165 real-world patient symptom queries paired with detailed physician responses across more than ten clinical specialties.

Retrieval Approaches. We compare four retrieval strategies that differ in their document and query representations.

BM25 (Robertson and Zaragoza, 2009) is a classic sparse probabilistic retrieval model that scores documents by the weighted overlap of query terms, applying a term-frequency saturation function and a document-length normalisation penalty. BM25 has long served as a strong baseline for ad-hoc retrieval and remains competitive with many neural approaches. We adopt BM25 parameters k_1=0.9 and b=0.4, and apply a title-boost factor of 2 by repeating title tokens at indexing time to approximate field-weighted BM25F scoring.

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	21.65	12.15	14.01	12.29	**16.75**	8.76	5.82	13.06
	BioASQ	**27.43**	12.54	**15.26**	**12.43**	16.11	**8.84**	**7.04**	**14.24**
	HealthCareMagic	19.90	**12.71**	14.41	11.87	16.22	8.76	6.43	12.90
	Medical Textbooks	22.07	12.54	15.04	12.30	15.47	**8.84**	6.30	13.22
	Yahoo Answers	20.05	12.48	15.04	11.47	15.87	8.72	6.03	12.81
LLaMA3.1-70B	w/o RAG	21.93	13.18	**17.19**	14.61	**18.35**	9.07	5.23	14.22
	BioASQ	**28.98**	13.10	15.95	14.03	15.71	9.13	5.73	**14.66**
	HealthCareMagic	23.58	13.35	16.28	**15.27**	17.63	**9.29**	5.70	14.44
	Medical Textbooks	24.71	13.14	15.73	14.96	15.50	9.21	5.72	14.14
	Yahoo Answers	23.68	**13.30**	16.18	14.78	16.22	9.26	**5.88**	14.19
Mistral-7B	w/o RAG	22.55	13.27	15.11	12.62	17.05	9.05	5.82	13.64
	BioASQ	**26.55**	13.49	15.50	12.78	16.81	9.39	6.57	**14.44**
	HealthCareMagic	24.09	**13.91**	15.93	12.78	**17.24**	9.46	6.42	14.26
	Medical Textbooks	22.81	13.45	15.30	**12.90**	16.45	9.28	6.41	13.80
	Yahoo Answers	23.96	13.68	**16.05**	13.15	17.21	**9.56**	**6.66**	14.32
Qwen2.5-7B	w/o RAG	20.95	12.36	15.13	12.24	15.80	8.47	5.41	12.91
	BioASQ	**23.93**	12.46	15.59	12.41	**16.29**	8.54	5.69	**13.56**
	HealthCareMagic	21.21	12.42	15.20	11.89	16.07	**8.58**	5.61	13.00
	Medical Textbooks	21.99	**12.49**	15.59	**12.42**	15.83	8.56	**5.87**	13.25
	Yahoo Answers	21.29	12.43	**15.74**	12.30	16.12	8.63	5.86	13.20
Qwen2.5-72B	w/o RAG	22.22	12.86	16.03	12.23	17.17	8.92	5.47	13.56
	BioASQ	**24.52**	12.83	15.83	12.21	17.38	8.97	5.61	**13.91**
	HealthCareMagic	23.60	12.90	15.68	11.76	17.67	**9.02**	5.43	13.72
	Medical Textbooks	23.69	**12.91**	15.92	**12.60**	17.32	8.98	5.61	13.86
	Yahoo Answers	23.60	12.89	**16.18**	12.10	**17.80**	**9.02**	**5.71**	13.90

Table 1: ROUGE-L by model and retrieval corpus (open-ended datasets).

TF-IDF (SPARCK JONES, 1972) represents both documents and queries as bag-of-words vectors weighted by term frequency–inverse document frequency, and ranks candidates by cosine similarity. Unlike BM25, TF-IDF applies no term-frequency saturation or document-length penalty, making it a simpler baseline for sparse lexical matching. We build a TF-IDF index with a vocabulary capped at 50,000 features and standard English stop-word removal.

MedCPT (Jin et al., 2023) is a biomedical dense retrieval model consisting of a *query encoder* and an *article encoder* trained contrastively on large-scale PubMed user search logs. Documents are encoded offline by the article encoder and stored as L2-normalised embeddings; at query time, the query encoder produces a query embedding and retrieval proceeds by maximum inner-product search. By capturing semantic similarity beyond exact term overlap, MedCPT is particularly well-suited to the biomedical domain.

Hybrid BM25 + MedCPT via RRF (Cormack et al., 2009) combines the BM25 and MedCPT ranked lists using Reciprocal Rank Fusion (RRF). Each document d ranked at position r in a ranked list receives a score $\frac{1}{k+r}$; with k=60, and the scores are summed across both lists. The final ranking is by descending combined RRF score. RRF is parameter-light and has been shown to consistently outperform individual rankers as well as more complex score-fusion methods (Cormack et al., 2009).

For all retrieval conditions, we retrieve the top k=5 documents and concatenate them as the retrieved context prepended to the generator prompt.

Implementation Details. All generation experiments are conducted with five open-source instruction-tuned models spanning two scales. The 7–8B models are **Qwen2.5-7B-Instruct** (Yang et al., 2025), **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024), and **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), each servable on a single GPU. The 70B-scale models are **LLaMA-3.1-70B-Instruct** (Grattafiori et al., 2024) and **Qwen2.5-72B-Instruct** (Yang et al., 2025), which serve as large-scale reference points to contextualise the small-model results.

All models are run in half-precision (FP16) with greedy decoding and a maximum of 300 newly generated tokens per response.

Experimental Setting. Each experimental condition is defined by a triple (*retriever*, *corpus*, *query dataset*). The retriever dimension covers five options: *No retrieval* (baseline), BM25, TF-IDF, MedCPT, and Hybrid (BM25 + MedCPT via RRF). The corpus dimension covers four knowledge bases: BioASQ/PubMed and Medical Textbooks as expert corpora, and Yahoo Answers and HealthCareMagic as layman corpora (for the base-

line condition both retriever and corpus are set to none). The query dimension covers the ten datasets described in Section 3, split evenly between layman (MeQSum, MedRedQA, MedicationQA, MASH-QA, ChatDoctor-iCliniq) and expert (BioASQ Task B, MedQuAD, MedQA-USMLE, MedMCQA, MMLU Medical) user types.

For the *w/o RAG* condition the model receives only the question in its prompt, with no retrieved context. For retrieval-augmented conditions, the top-k retrieved passages are prepended to the question in a fixed prompt template. Each condition is run independently for every generator model, and all per-dataset query pools described in Section 3 are also available for few-shot prompting ablations. The complete set of conditions spans every combination of retriever, corpus, and query dataset, yielding a large-scale cross-model, cross-retriever, cross-dataset evaluation.

4 Results

We present results separately for open-ended QA, evaluated with ROUGE-L as the primary metric (ROUGE-1, ROUGE-2, METEOR, BLEU, and BERTScore in Appendix C), and for multiple-choice QA, evaluated with accuracy.

Open-ended QA. Table 1 reports ROUGE-L across seven open-ended datasets (five layman and two expert), averaged over all retrieval conditions per corpus. Across all models, retrieval yields small and inconsistent improvements over the no-retrieval baseline. The largest gains appear on the BioASQ open-ended task, where the BioASQ/PubMed corpus consistently provides the strongest lift: for example, LLaMA-3.1-8B improves from 21.65 to 27.43 ROUGE-L. However, for the remaining six datasets, changes from the baseline are typically under 1 ROUGE-L point and often negative. Averaged across all seven datasets, the maximum retrieval benefit over no-retrieval is 1.18 points (LLaMA-3.1-8B: 13.06 baseline vs. 14.24 with BioASQ); for all other models the gain is smaller still. ROUGE-1, ROUGE-2, METEOR, BLEU, and BERTScore results (Appendix C) show the same pattern.

A consistent observation is that backbone model choice matters far more than retrieval configuration. Mistral-7B lags behind both LLaMA-3.1-8B and Qwen2.5-7B regardless of the retrieval setup, and the 70B-scale models (LLaMA-3.1-70B, Qwen2.5-72B) are consistently stronger than all 7–8B variants. The gap between any two retrieval conditions for the same model is almost always smaller than the gap between two different backbone models using the same conditions. Expert and layman retrieval corpora produce similar results in most open-ended settings, differing by less than 1 ROUGE-L point on average.

Model	Data	MCQA	MQA	MMLU	Avg
LLaMA3.1-8B	w/o RAG	**80.8**	83.8	**83.7**	**82.8**
	BioASQ	75.9	**84.6**	82.3	80.9
	HCM	73.7	84.2	74.0	77.3
	Textbook	74.1	84.1	83.3	80.5
	Yahoo	74.6	83.7	81.5	79.9
LLaMA3.1-70B	w/o RAG	**81.0**	89.1	89.2	86.4
	BioASQ	79.8	90.8	**90.2**	**86.9**
	HCM	78.9	80.5	87.5	82.3
	Textbook	80.2	81.1	89.5	83.6
	Yahoo	79.1	**91.7**	88.7	86.5
Mistral-7B	w/o RAG	**72.6**	**77.9**	**76.7**	**75.7**
	BioASQ	61.3	72.4	72.2	68.6
	HCM	63.3	73.2	72.0	69.5
	Textbook	66.0	73.7	77.3	72.3
	Yahoo	65.5	74.0	74.2	71.2
Qwen2.5-7B	w/o RAG	**80.0**	**83.7**	**86.3**	**83.3**
	BioASQ	75.8	81.1	82.2	79.7
	HCM	74.9	81.4	82.8	79.7
	Textbook	76.6	81.1	85.7	81.1
	Yahoo	77.5	81.4	85.2	81.4
Qwen2.5-72B	w/o RAG	**82.5**	81.9	**92.5**	**85.6**
	BioASQ	80.1	81.6	91.2	84.3
	HCM	77.3	**85.8**	90.7	84.6
	Textbook	80.1	83.5	91.2	84.9
	Yahoo	79.8	81.8	90.5	84.0

Table 2: Accuracy by model and retrieval corpus. MCQA denotes MedMCQA, MQA denotes MedQA, and HCM denotes HealthCareMagic.

Multiple-choice QA. Table 2 reports accuracy grouped by dataset subset (MCQA: MedQA + MedMCQA; QA: open-ended expert; MMLU: six MMLU medical subjects). For smaller models (LLaMA-3.1-8B, Mistral-7B, Qwen2.5-7B), retrieval frequently *hurts* accuracy relative to the no-retrieval baseline. Mistral-7B drops from 75.7 to 68.6–72.3 across all retrieval corpora. The larger models (LLaMA-3.1-70B, Qwen2.5-72B) are more robust, maintaining accuracy within 1–2 points of the baseline across all conditions, but still show no consistent gain. As with the open-ended setting, backbone choice dominates: Qwen2.5-72B's no-retrieval accuracy of 85.6 exceeds the best retrieval configuration of any 7B model by over 2 points.

Effect of Retrieval Method. Table 3 reports accuracy averaged across the three close-ended datasets from Table 2 (MedMCQA, MedQA-USMLE, and MMLU Medical), broken down by

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	79.34	79.57	**80.94**	80.65
	HealthCareMagic	72.62	74.59	74.88	**76.97**
	Medical Textbooks	79.47	**80.34**	79.53	79.08
	Yahoo Answers	79.27	**79.74**	77.91	77.14
	Average	77.68	**78.56**	78.32	78.46
LLaMA3.1-70B	BioASQ	85.56	82.11	**86.52**	85.06
	HealthCareMagic	81.92	80.99	81.52	**82.14**
	Medical Textbooks	81.54	82.87	82.68	**83.29**
	Yahoo Answers	85.51	81.69	81.32	**85.77**
	Average	83.64	81.91	83.01	**84.06**
Mistral-7B	BioASQ	66.82	68.13	**68.52**	67.73
	HealthCareMagic	68.36	68.81	**69.12**	68.64
	Medical Textbooks	70.66	**71.47**	70.52	69.37
	Yahoo Answers	68.59	69.43	69.72	**70.94**
	Average	68.61	69.46	**69.47**	69.17
Qwen2.5-7B	BioASQ	78.37	78.88	**79.61**	79.49
	HealthCareMagic	78.39	78.55	77.87	**78.84**
	Medical Textbooks	**80.28**	80.15	**80.28**	79.39
	Yahoo Answers	80.52	80.03	**80.81**	79.69
	Average	79.39	79.40	**79.64**	79.35
Qwen2.5-72B	BioASQ	83.05	**83.86**	83.58	83.16
	HealthCareMagic	82.34	82.70	82.80	**83.84**
	Medical Textbooks	**84.72**	84.37	83.04	82.63
	Yahoo Answers	83.22	83.44	**83.69**	82.90
	Average	83.33	**83.59**	83.28	83.13

Table 3: Accuracy by retrieval method (close-ended datasets).

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	13.93	**14.10**	13.57	13.17
	HealthCareMagic	12.47	**12.80**	12.58	12.71
	Medical Textbooks	12.90	13.01	**13.08**	12.93
	Yahoo Answers	12.59	**12.73**	12.58	12.23
	Average	12.98	**13.16**	12.95	12.76
LLaMA3.1-70B	BioASQ	13.95	**14.58**	13.78	13.59
	HealthCareMagic	14.19	14.31	**14.39**	14.21
	Medical Textbooks	13.57	13.83	**14.00**	13.85
	Yahoo Answers	13.68	**13.97**	**13.97**	13.41
	Average	13.85	**14.17**	14.04	13.77
Mistral-7B	BioASQ	14.19	**14.34**	13.91	13.92
	HealthCareMagic	14.08	14.11	14.01	**14.14**
	Medical Textbooks	13.71	**13.74**	13.59	13.67
	Yahoo Answers	14.14	14.16	14.13	**14.20**
	Average	14.03	**14.09**	13.91	13.98
Qwen2.5-7B	BioASQ	13.26	**13.49**	13.23	12.97
	HealthCareMagic	12.80	**12.90**	12.87	**12.90**
	Medical Textbooks	12.98	**13.22**	13.04	13.06
	Yahoo Answers	12.99	**13.13**	**13.13**	13.05
	Average	13.01	**13.19**	13.07	13.00
Qwen2.5-72B	BioASQ	13.80	**13.82**	13.73	13.75
	HealthCareMagic	13.57	13.56	13.65	**13.70**
	Medical Textbooks	**13.81**	13.74	13.61	13.77
	Yahoo Answers	13.63	13.72	13.84	**13.85**
	Average	13.70	13.71	13.71	**13.77**

Table 4: ROUGE-L by retrieval method (open-ended datasets).

retrieval method rather than corpus. Table 4 reports ROUGE-L averaged across the seven open-ended datasets from Table 1 (BioASQ, ChatDoctor-iCliniq, MashQA, MedicationQA, MedQuAD, MedRedQA, and MeQSum), again broken down by retrieval method. Together, these two tables allow direct comparison of BM25, Hybrid (RRF), MedCPT, and TF-IDF across question types and retrieval corpora. Differences among methods are within 1–2 points for any model–corpus combination. The Hybrid retriever shows marginal advantages in several configurations, but no method consistently dominates. MedCPT, despite domain-specific training, does not systematically outperform lexical BM25. Full per-metric breakdowns by retriever type (ROUGE-1, ROUGE-2, METEOR, BLEU, BERTScore) are in Appendix C.

5 Ablation Study

We conduct two ablations to understand how retrieval depth and few-shot context affect performance. Both use a stratified subset of the test queries with BM25 retrieval from BioASQ. Additional open-ended metric trends are visualised in Appendix D (Figures 7 and 8).

Number of Retrieved Documents (Top-k). Figures 5 and 6 show accuracy and ROUGE-L as k varies over $\{1, 3, 5, 10, 25, 50\}$. For open-ended metrics, performance reaches a plateau by k=5:

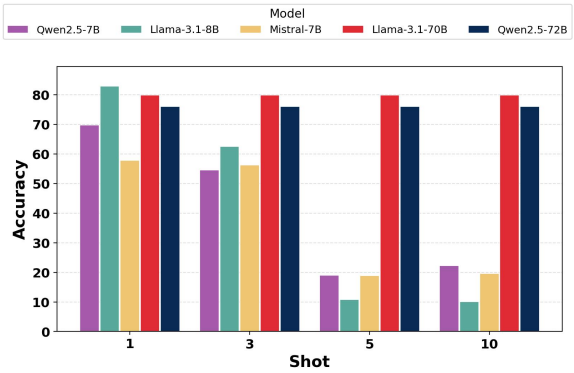


Figure 3: Close-ended accuracy across shot counts (1, 3, 5, 10).

ROUGE-L changes by less than 0.2 points between k=5 and k=50 for all models, indicating that additional retrieved documents add no useful signal once the context budget is satisfied. For close-ended accuracy the picture is less uniform: LLaMA-3.1-8B peaks at k=5 (72.83%) before declining, while Qwen2.5-7B and LLaMA-3.1-70B reach their best performance at $k \geq 25$. Mistral-7B declines steadily after k=3, reaching 51.22% at $k \geq 25$. These results confirm that k=5 is a reasonable default: it matches or closely approaches the optimum for most models while keeping context length manageable. Additional open-ended metric trends across all k values are shown in Figure 8 in the appendix.

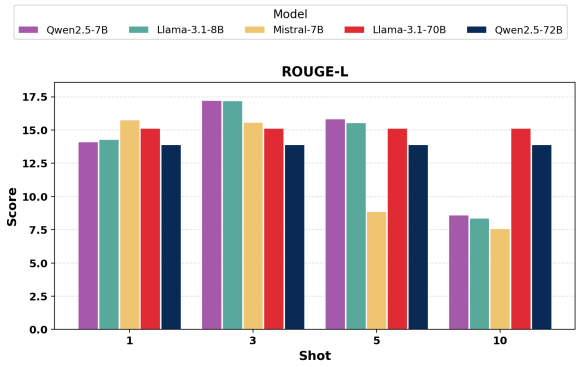


Figure 4: Open-ended ROUGE-L across shot counts (1, 3, 5, 10).

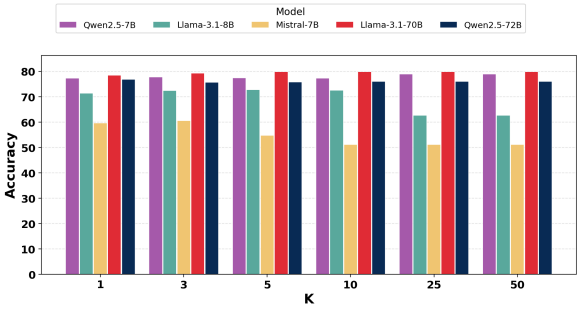


Figure 5: Close-ended accuracy across top-k (1, 3, 5, 10, 25, 50).

Few-shot Prompting. Figures 3 and 4 show accuracy and ROUGE-L as the number of in-context examples varies over $\{1, 3, 5, 10\}$. Larger models (LLaMA-3.1-70B, Qwen2.5-72B) are essentially unaffected by shot count across all metrics, suggesting they can extract the task pattern from a single example or from zero-shot prompting equally well. In contrast, smaller 7–8B models show sharp degradation at 5 and 10 shots: LLaMA-3.1-8B accuracy collapses from 82.89% (1-shot) to 10.06% (10-shot), and ROUGE-L drops from 14.29 to 8.38, as the long few-shot context overwhelms the model's ability to locate the target instruction. Mistral-7B and Qwen2.5-7B follow the same pattern. Notably, 3-shot prompting is the sweet spot for open-ended ROUGE-L: LLaMA-3.1-8B reaches 17.19 at 3 shots (vs. 14.29 at 1-shot), and Qwen2.5-7B reaches 17.22, before degrading at higher shot counts. For MCQ accuracy, even 3 shots already reduces performance for most small models, pointing to the inherent tension between providing helpful demonstrations and staying within the model's effective context capacity. Additional open-ended metric trends across all shot counts are shown in Figure 7 in the appendix.

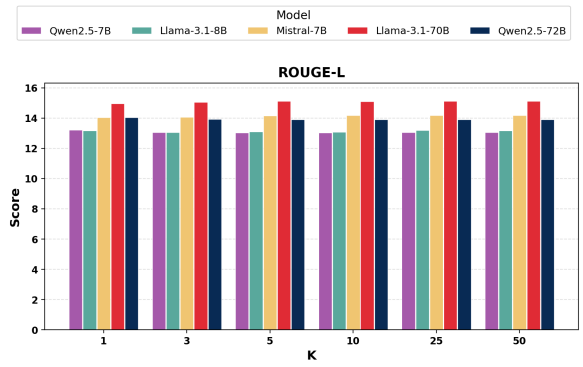


Figure 6: Open-ended ROUGE-L across top-k values.

Model	w/o RAG	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	0.410	0.580	0.540	0.440	0.530
LLaMA3.1-70B	0.410	0.660	0.610	0.540	0.660
Mistral-7B	0.460	0.510	0.490	0.390	0.530
Qwen2.5-7B	0.410	0.380	0.400	0.360	0.390
Qwen2.5-72B	0.380	0.350	0.350	0.360	0.380

Table 5: Accuracy of LLMs across retrieval methods in the oracle retrieval setting, where all retrieved documents are relevant (clean context).

Quality of the Retrieval Analysis. Tables 5 and 6 show two important problems for retrieval-augmented generation in the biomedical domain. For this analysis, we use the BioASQ corpus as the retrieval source and evaluate on PubMedQA (Jin et al., 2019), a benchmark of expert-annotated yes/no/maybe biomedical research questions derived from PubMed abstracts. Since both the retrieval corpus and evaluation dataset come from PubMed, this provides a controlled setting for studying whether retrieved biomedical papers help models answer research questions. To evaluate retrieval quality, we use an LLM-as-a-judge framework to determine whether the retrieved context contains enough information to answer the question correctly. We then select 100 questions where all retrieval methods retrieved context judged to be relevant. The questions are the same across all retrieval methods, but the retrieved documents can differ depending on the retriever.

Table 5 shows that even when all retrieved contexts contain the correct information, retrieval only leads to limited and inconsistent improvements. For example, LLaMA3.1-70B improves substantially with BM25 retrieval (0.410 $\rightarrow$ 0.660), while Qwen2.5-72B shows almost no improvement across retrieval methods. In several cases, simple sparse retrieval methods such as BM25 and TF-IDF perform better than MedCPT. These results suggest that retrieving relevant evidence alone is not

Model	w/o RAG	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	0.410	0.300	0.250	0.300	0.240
LLaMA3.1-70B	0.410	0.260	0.260	0.340	0.290
Mistral-7B	0.460	0.310	0.330	0.350	0.340
Qwen2.5-7B	0.410	0.310	0.260	0.290	0.250
Qwen2.5-72B	0.380	0.250	0.240	0.280	0.230

Table 6: Accuracy of LLMs across retrieval methods in the noisy retrieval setting, where 20 unrelated documents are mixed with retrieved results (distracted context).

enough to guarantee better performance. Instead, many models still struggle to correctly use and reason over the retrieved information. Table 6 further shows that current models are highly sensitive to irrelevant context. When we add 20 unrelated documents to the retrieved evidence, performance drops substantially across nearly all models and retrieval methods. For example, LLaMA3.1-70B decreases from 0.660 to 0.260 under BM25 retrieval, while Mistral-7B drops from 0.530 to 0.340 under TF-IDF retrieval. In many cases, performance becomes worse than using no retrieval at all. Overall, these results show that current biomedical RAG systems remain brittle. Even when relevant evidence is retrieved successfully, small amounts of distracting context can strongly reduce answer accuracy.

Implications. Our results suggest a more cautious view of biomedical RAG. Retrieval can help, but only when the system retrieves information that is actually relevant to the question. This is not guaranteed, especially when the answer is absent from the corpus or when the retrieved passages are only loosely related. In these cases, retrieval may add little useful information and can introduce misleading context.

Even when relevant evidence is retrieved, the model still has to understand and use it correctly. Our clean retrieval analysis shows that relevant context does not always improve performance, suggesting that evidence use is a major bottleneck. The noisy retrieval results make this concern stronger: adding unrelated documents to useful evidence often hurts performance, sometimes making RAG worse than no retrieval at all. Future biomedical RAG systems, therefore, need better evidence filtering, reranking, and generation methods that can identify useful passages while ignoring distractors.

6 Conclusion

We presented a large-scale evaluation of retrieval-augmented generation for biomedical question answering using five open-weight, instruction-tuned models ranging from 7B to 72B parameters. Across all five models, ten datasets, four retrieval methods, and four retrieval corpora, retrieval yields only small and inconsistent improvements over a no-retrieval baseline, typically within 1–2 points on any metric. In contrast, backbone model choice has a substantially larger effect: the gap between a 7B model and its 70B counterpart often exceeds the gain from any retrieval configuration. Expert and layman retrieval corpora also perform similarly in most settings, and differences across retrieval methods (BM25, TF-IDF, MedCPT, Hybrid) remain minor throughout.

Our ablation studies further support this overall pattern. Increasing the number of retrieved documents beyond $k{=}5$ provides little additional benefit for open-ended settings, and few-shot prompting yields a modest gain at 3 shots for smaller models but degrades sharply at higher counts, with small models struggling under longer few-shot contexts. Larger models are comparatively stable across shot counts, but they also show limited benefit from retrieval augmentation.

Taken together, these findings suggest that improving retrieval quality alone may not be sufficient to substantially improve biomedical QA performance in these settings. One possible explanation is that current models, especially smaller ones, do not consistently make effective use of retrieved evidence, though our experiments do not directly measure evidence utilization or grounding. This points to several directions for future work, including training or fine-tuning methods that better support evidence integration, post-retrieval reranking or filtering to reduce context noise, and evaluation frameworks that more directly assess faithfulness and grounding rather than relying only on reference-based metrics. More broadly, an important open question is when retrieval is actually necessary, and whether we can better identify cases where the required knowledge is already contained within the model.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

Limitations

Our study has several limitations. First, we evaluate retrieval-augmented generation using only reference-based downstream metrics such as ROUGE-L, BLEU, METEOR, BERTScore, and accuracy, rather than direct measures of faithfulness or evidence grounding. For example, a model may produce a correct answer from its parametric knowledge without actually using the retrieved documents, or it may copy surface details from retrieval without truly improving medical correctness. This limitation is not major for our study because our main goal is a comparative evaluation across models, retrievers, corpora, and no-retrieval baselines within a single, consistent framework, and these metrics are sufficient to support our central finding that retrieval provides only small and inconsistent gains.

Second, our experiments are limited to five open-weight instruction-tuned models and do not include proprietary frontier systems such as GPT-4-class medical assistants. It is possible that stronger closed models use retrieved evidence more effectively, especially in cases requiring multi-step reasoning over documents. This limitation is not major for our study because our paper is specifically motivated by practical, deployable biomedical QA settings, where open-weight 7B–72B models are realistic choices, and we also include both small and large open models to test whether the observed pattern holds across scales.

Third, our retrieval setup uses a fixed top-k pipeline with four retrievers and four corpora, but does not explore more complex retrieval strategies such as adaptive retrieval, document re-ranking, iterative retrieval, or task-specific chunking. For instance, some questions may require retrieving fewer but more precise passages, while others may benefit from multi-hop retrieval or filtering noisy evidence before generation. This limitation is not major for our study because we intentionally focus on strong, standard retrieval baselines widely used in prior biomedical RAG work, which makes the comparison clean and allows us to show that, even with several commonly used retrieval choices, the gains remain modest.

Finally, our evaluation mixes expert and layman biomedical QA datasets, but the study does not separately analyze all possible sources of variation across question type, answer length, or knowledge intensity. For example, retrieval may be more useful for highly specialized factoid questions than for common consumer-health questions that models may already answer from pretraining alone. This limitation is not major for our study because the breadth of datasets is a strength of the paper overall: the consistency of the pattern across ten datasets suggests that the weak benefit of retrieval is not tied to a single benchmark or user population.

References

A. B. Abacha, Y. Mrabet, M. Sharp, T. R. Goodwin, S. E. Shooshan, and D. Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. *Stud Health Technol Inform*, 264:25–29. 1879-8365 Abacha, Asma Ben Mrabet, Yassine Sharp, Mark Goodwin, Travis R Shooshan, Sonya E Demner-Fushman, Dina Journal Article Netherlands 2019/08/24 Stud Health Technol Inform. 2019 Aug 21;264:25-29. doi: 10.3233/SHTI190176.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Asma Ben Abacha and Dina Demner-Fushman. 2019a. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019b. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

A. Krithara, A. Nentidis, K. Bougiatiotis, and G. Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Sci Data*, 10(1):170. 2052-4463 Krithara, Anastasia Orcid: 0000-0003-0491-4507 Nentidis, Anastasios Bougiatiotis, Konstantinos Paliouras, Georgios Dataset Journal Article England 2023/03/28 Sci Data. 2023 Mar 27;10(1):170. doi: 10.1038/s41597-023-02068-4.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, Grigorios Tsoumakas, George Giannakoulas, Alexandra Bekiaridou, Athanasios Samaras, Giorgio Maria Di Nunzio, Nicola Ferro, Stefano Marchesin, Marco Martinelli, Gianmaria Silvello, and Georgios Paliouras. 2025. Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Cham. Springer Nature Switzerland, Springer Nature Switzerland.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

KAREN SPARCK JONES. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Yahoo! Research. 2009. Yahoo! Webscope Datasets Catalog. Technical report, Yahoo! Inc. 19 Datasets Available. Accessed via Stanford InfoLab.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ming Zhang, Shuai Dou, Ziyang Wang, and Yunfang Wu. 2022. Focus-driven contrastive learning for medical question summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6176–6186, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Datasets

Our experiments use ten biomedical and consumer-health query datasets and four retrieval corpora. Table 7 summarises each evaluation dataset, its source size, and the number of examples selected for evaluation; Table 8 describes the knowledge bases indexed for retrieval. Full details on dataset splits and the few-shot query pools are provided in Section 3.

Datasets are grouped by intended user type: *layman* datasets reflect consumer health inquiries in everyday language, while *expert* datasets target biomedical professionals or medical students. The *Query Set* column reports the eligible pool size used for few-shot prompting; the *Test Set* column reports the number of examples used as evaluation queries. All corpora in Table 8 are indexed as whole records without further chunking; for Q&A-format corpora (Yahoo Answers and HealthCareMagic) each document concatenates the question or title with the answer body.

B Prompt Templates

All generation experiments use two system prompts and two user message templates, combined via each model's native chat template.

Layman system prompt. Used when the query dataset belongs to the *layman* user type (MeQSum, MedRedQA, MedicationQA, MASH-QA, ChatDoctor-iCliniq):

Layman System Prompt

You are a helpful health assistant answering questions from members of the general public. Use simple, everyday language that a non-medical person can easily understand. Avoid medical jargon. Be clear, friendly, and concise.

User Type	Dataset	Dataset Type	Dataset Content	Query Set	Test Set
Layman	MeQSum (Ben Abacha and Demner-Fushman, 2019a)	Open-ended Q&A	1,000 consumer health questions from the U.S. National Library of Medicine, manually summarized by medical experts (inter-annotator agreement: 96.9%)	500	500
	MedRedQA (Nguyen et al., 2023)	Open-ended Q&A	51,000 consumer question–physician answer pairs from Reddit `/r/AskDocs` (2013–2022); answers from verified doctors only; ~1k entries enriched with PubMed evidence	45,863	1,000
	MedicationQA (Abacha et al., 2019)	Open-ended Q&A	690 real consumer medication questions annotated with drug focus and question type (dosage, side effects, interactions, etc.); answers sourced from MedlinePlus, DailyMed, FDA, and Mayo Clinic	189	500
	MASH-QA (Zhu et al., 2020)	Open-ended Q&A	34,808 consumer health Q&A pairs from WebMD; extractive, multi-span answers (avg. 67 words) curated by healthcare experts from articles averaging 696 words	19,989	1,000
	ChatDoctor-iCliniq (Li et al., 2023)	Open-ended Q&A	7,321 real patient–physician conversations from iCliniq.com spanning infectious disease, dermatology, cardiology, neurology, and other specialties	6,321	1,000
Expert	BioASQ Task B (Nentidis et al., 2025)	Summary Q&A	Expert-curated biomedical questions paired with PubMed-grounded answers; covers yes/no, factoid, list, and summary types with gold concepts, snippets, and RDF triples	1,283	80
	MedQuAD (Ben Abacha and Demner-Fushman, 2019b)	Open-ended Q&A	47,457 Q&A pairs from 12 NIH websites covering 37 question types (treatment, diagnosis, side effects, etc.) across diseases, drugs, and medical tests; enriched with UMLS CUI metadata	15,407	1,000
	MedQA-USMLE (Jin et al., 2021)	Multiple-Choice Q&A	~11,500 clinical vignette MCQs from the USMLE (Steps 1–3); each vignette describes patient demographics, symptoms, and history followed by a 4-option diagnostic or management question	10,178	1,273
	MedMCQA (Pal et al., 2022)	Multiple-Choice Q&A	194k+ MCQs from AIIMS & NEET PG medical entrance exams spanning 21 subjects and 2,400+ topics (anatomy, pharmacology, pathology, surgery, psychiatry, etc.); includes answer explanations	182,822	1,000
	MMLU Medical (Hendrycks et al., 2021)	Multiple-Choice Q&A	1,242 four-choice questions from six MMLU medical subjects (clinical knowledge, medical genetics, anatomy, etc.), drawn from GRE and USMLE practice exams; evaluated in zero/few-shot settings	642	600

Table 7: Evaluation query datasets grouped by user type. *Layman*: consumer health queries in everyday language; *Expert*: biomedical/clinical questions. "Query Set": few-shot pool size; "Test Set": evaluation examples used.

Expert system prompt. Used for *expert* datasets (BioASQ Task B, MedQuAD, MedQA-USMLE, MedMCQA, MMLU Medical):

Expert System Prompt

You are a clinical decision support assistant. Answer questions from healthcare professionals using precise medical terminology. Provide evidence-based, clinically detailed responses with relevant diagnostic and therapeutic considerations.

User message (without retrieval)

User Message

```
QUESTION: {query}
ANSWER:
```

User message (with retrieval)

User Message

```
Use the following retrieved passages to
help answer the question.
RETRIEVED CONTEXT:
{context}
QUESTION: {query}
ANSWER:
```

User Type	Dataset	Source Content	Source Rows	Documents Retained	Dataset Link
Expert	BioASQ Task A / PubMed (Krithara et al., 2023)	PubMed abstracts with human-assigned MeSH annotations (avg. 12.68 terms per article; up to 29,681 distinct MeSH terms across 16.2M articles)	16,218,838	16,218,838	Link
	Medical textbooks (Xiong et al., 2024)	Retrieval-friendly chunks ($\leq$1,000 chars each) drawn from 18 authoritative biomedical textbooks spanning anatomy, physiology, pharmacology, pathology, and clinical medicine	125,847	125,847	Link
Layman	Yahoo Answers (Yahoo! Research, 2009)	Open-domain community Q&A posts with best-answer labels and topic categories; user identities fully anonymized	1,400,000	1,238,506	Link
	HealthCareMagic (Li et al., 2023)	Real-world patient symptom queries paired with detailed physician responses (diagnosis, treatment, and referral advice) across 10+ clinical specialties	112,165	112,165	Link

Table 8: Retrieval corpora grouped by user type. *Expert*: technical biomedical sources; *Layman*: community health and general Q&A. "Documents Retained": records after quality filtering.

where `{context}` is a concatenation of the top-k retrieved passages, each formatted as `[Passage N (source: {source})]: {text}`. The final prompt is produced by wrapping these system and user messages in each model's chat template.

C Full Results by Metric

This section provides per-dataset performance tables for all evaluated metrics beyond ROUGE-L (reported in the main paper). Tables 9–13 report ROUGE-2, ROUGE-1, BERTScore, METEOR, and BLEU respectively, broken down by model and retrieval corpus across the seven open-ended datasets. Tables 14–18 further break down ROUGE-1, ROUGE-2, BLEU, METEOR, and BERTScore by retrieval method (BM25, Hybrid, MedCPT, TF-IDF) across all four corpora.

ROUGE-2 (Table 9). The pattern mirrors ROUGE-L: the BioASQ/PubMed corpus produces the largest gains, and only on the BioASQ expert open-ended task. For example, LLaMA-3.1-8B improves from 14.03 to 18.34, and LLaMA-3.1-70B from 14.43 to 19.25, while all other datasets see gains under 1 point or negative effects. The average improvement over baseline is at most 0.53 points (LLaMA-3.1-8B: 5.82 $\rightarrow$ 6.35), and for Qwen2.5-72B Yahoo Answers produces the best average (6.39), marginally ahead of BioASQ (6.37), illustrating how small these corpus-level differences are.

ROUGE-1 (Table 10). Again, the BioASQ corpus helps on the BioASQ dataset (gains of 3–5 points for all models) while effects on lay datasets are within $\pm$1 point. Mistral-7B shows an above-average improvement with BioASQ corpus on the BioASQ open-ended task (37.56 $\rightarrow$ 40.46 with BioASQ corpus; 40.53 with Qwen2.5-72B), confirming domain-matched retrieval has local benefit. Averaged over all datasets the maximum gain is 0.65 points (LLaMA-3.1-8B baseline 22.49 $\rightarrow$ best 22.88).

BERTScore (Table 11). BERTScore is notably more stable than any ROUGE metric: the gap between the no-retrieval baseline and the best retrieval condition is under 0.7 points for all models. For instance, LLaMA-3.1-8B moves from 52.47 (baseline) to at best 52.85 (BioASQ corpus), a gain of just 0.38 points. This suggests that while retrieved context can slightly shift surface n-gram overlap, the overall semantic content of model outputs barely changes, consistent with the view that 7–8B models are not effectively incorporating the retrieved evidence.

METEOR (Table 12). METEOR shows small, mixed effects: the BioASQ corpus provides a modest boost on the BioASQ dataset (e.g., LLaMA-3.1-8B: 29.84 $\rightarrow$ 30.92; Qwen2.5-72B: 31.11 $\rightarrow$ 34.05), but on lay datasets retrieval often slightly lowers METEOR, particularly for the 70B models where the baseline exceeds all retrieval conditions on several tasks (e.g., LLaMA-3.1-70B average: 18.92 baseline vs. 17.03–17.12 across all corpora).

BLEU (Table 13). BLEU scores are generally very low for lay datasets ($<$ 2 across all conditions), underscoring that n-gram precision is a weak signal for open-ended health QA. The BioASQ corpus produces notable gains on the expert BioASQ dataset (LLaMA-3.1-8B: 12.92 $\rightarrow$ 19.08; LLaMA-3.1-70B: 13.32 $\rightarrow$ 19.76), but all other datasets improve by less than 0.3 BLEU points, and many worsen. Average BLEU across all datasets improves by 0.82 points at most.

Retrieval method breakdown (Tables 14–18). Across all five metrics, differences among BM25, Hybrid (RRF), MedCPT, and TF-IDF are consistently within 0.5 metric points for any model–corpus combination. The Hybrid retriever shows a slight edge in several configurations (particularly ROUGE-L and ROUGE-1), while TF-IDF is competitive with BM25 despite its greater simplicity. No single retrieval method dominates across all metrics and models, reinforcing the conclusion that retrieval architecture choice is secondary to corpus and model selection.

D Ablation Study: Additional Figures

This section provides additional figures for the two ablation studies described in Section 5. Figure 7 shows BERTScore, METEOR, BLEU, ROUGE-2, and ROUGE-1 trends under few-shot prompting across all five models on open-ended questions. Figure 8 shows the same metrics across top-k values.

Few-shot: additional metrics (Figure 7). All six open-ended metrics tell a consistent story. For the larger models (LLaMA-3.1-70B and Qwen2.5-72B), all metrics are flat across all shot counts: for example, METEOR stays at 17.44–17.45 for LLaMA-3.1-70B and BERTScore stays at 52.69 for Qwen2.5-72B regardless of shot count. For smaller models, the 3-shot sweet spot and subsequent collapse are visible in every metric. Specifically, ROUGE-1 peaks at 3 shots for LLaMA-3.1-8B (26.29) and Qwen2.5-7B (26.61) before collapsing to 12.19 and 13.52 at 10 shots. METEOR follows the same pattern: LLaMA-3.1-8B peaks at 19.99 (3 shots) vs. 9.12 (10 shots), and Qwen2.5-7B at 20.54 (3 shots) vs. 10.43 (10 shots). BERTScore shows a more severe drop for Mistral-7B: from 54.32 at 1 shot to 40.96 at 10 shots, a 13-point collapse. BLEU, while numerically small, also collapses dramatically, LLaMA-3.1-8B drops from 5.15 (3 shots) to 1.15 (10 shots), and Mistral-7B from 4.66 (3 shots) to 0.88 (10 shots), confirming that higher shot counts severely degrade output quality at this scale.

Top-k: additional metrics (Figure 8). The plateau behavior seen in ROUGE-L (main paper) extends to all six additional metrics. ROUGE-1 changes by at most 0.13 points from k=5 to k=50 across all models. ROUGE-2 is similarly stable: for example, LLaMA-3.1-70B moves from 7.44 (k=5) to 7.45 (k=50), a negligible change. METEOR plateaus by k=5 for most models, with variations under 0.1 between k=5 and k=50. BERTScore is the most stable metric of all: for LLaMA-3.1-8B, it ranges only from 51.33 (k=1) to 51.53 (k=10), a 0.20-point spread across all six k values. The transition from k=1 to k=5 accounts for nearly all the variation, and additional passages beyond k=5 provide no measurable benefit in any metric for any model.

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	14.03	3.26	6.27	4.26	8.36	2.44	2.10	5.82
	BioASQ	18.34	3.51	5.81	3.82	7.81	2.53	2.63	6.35
	HealthCareMagic	12.97	3.69	5.64	3.72	8.10	2.51	2.23	5.55
	Medical Textbooks	13.39	3.51	5.74	3.71	7.18	2.50	2.30	5.48
	Yahoo Answers	12.37	3.48	5.90	3.57	7.76	2.48	2.20	5.39
LLaMA3.1-70B	w/o RAG	14.43	3.55	7.41	5.42	9.32	2.40	1.88	6.34
	BioASQ	19.25	3.51	5.94	4.95	7.90	2.28	2.21	6.58
	HealthCareMagic	15.52	3.66	6.49	5.89	9.29	2.27	2.15	6.47
	Medical Textbooks	15.86	3.56	5.94	5.47	7.82	2.31	2.19	6.16
	Yahoo Answers	15.31	3.58	6.28	5.39	8.45	2.28	2.23	6.22
Mistral-7B	w/o RAG	14.18	3.54	6.48	4.03	7.96	2.53	2.09	5.83
	BioASQ	17.02	3.58	5.89	3.78	7.49	2.62	2.42	6.11
	HealthCareMagic	14.47	3.94	6.34	3.86	8.23	2.64	2.46	5.99
	Medical Textbooks	13.80	3.63	5.87	4.05	7.25	2.56	2.40	5.65
	Yahoo Answers	14.75	3.64	6.43	4.02	8.19	2.67	2.47	6.02
Qwen2.5-7B	w/o RAG	13.42	3.30	6.70	3.98	7.20	2.44	1.83	5.55
	BioASQ	15.68	3.38	6.50	3.77	7.19	2.48	2.13	5.88
	HealthCareMagic	13.32	3.37	6.27	3.50	7.43	2.49	1.96	5.48
	Medical Textbooks	14.02	3.44	6.40	3.87	6.95	2.45	2.14	5.61
	Yahoo Answers	13.48	3.37	6.47	3.79	7.45	2.49	2.19	5.61
Qwen2.5-72B	w/o RAG	14.60	3.49	7.10	3.94	8.64	2.79	1.86	6.06
	BioASQ	16.73	3.67	7.03	3.85	8.54	2.81	1.99	6.37
	HealthCareMagic	16.03	3.70	6.93	3.72	9.11	2.83	1.88	6.31
	Medical Textbooks	15.84	3.68	7.13	4.21	8.49	2.77	2.00	6.30
	Yahoo Answers	16.27	3.54	7.06	3.79	9.25	2.83	2.02	6.39

Table 9: ROUGE-2 by model and retrieval corpus (open-ended datasets).

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	36.34	22.91	24.67	20.15	30.51	15.46	7.36	22.49
	BioASQ	40.24	23.93	25.59	18.91	27.02	15.61	8.84	22.88
	HealthCareMagic	32.93	24.57	24.44	18.60	29.30	15.52	8.25	21.94
	Medical Textbooks	34.70	24.06	25.33	18.93	26.58	15.54	8.06	21.89
	Yahoo Answers	33.68	23.93	25.32	18.17	28.70	15.39	7.61	21.83
LLaMA3.1-70B	w/o RAG	36.80	24.87	28.54	22.23	31.57	15.73	6.76	23.79
	BioASQ	41.45	24.77	24.03	20.40	24.68	15.60	7.27	22.60
	HealthCareMagic	37.14	24.93	24.74	22.31	28.36	15.78	7.29	22.94
	Medical Textbooks	37.13	24.84	23.85	21.65	23.59	15.70	7.27	22.00
	Yahoo Answers	36.85	24.99	24.00	21.25	25.75	15.79	7.52	22.31
Mistral-7B	w/o RAG	37.56	25.83	27.02	20.79	31.15	16.38	7.70	23.78
	BioASQ	40.46	26.16	26.82	20.35	28.71	16.94	8.66	24.01
	HealthCareMagic	38.66	26.97	27.78	20.50	30.78	16.99	8.57	24.32
	Medical Textbooks	36.24	25.98	26.62	21.03	29.13	16.73	8.46	23.46
	Yahoo Answers	37.99	26.58	27.77	20.63	30.59	17.17	8.91	24.23
Qwen2.5-7B	w/o RAG	36.89	24.77	27.60	20.11	29.91	15.82	7.56	23.24
	BioASQ	38.95	24.79	27.49	20.09	29.51	15.76	7.71	23.47
	HealthCareMagic	36.58	24.67	27.22	19.21	30.42	15.78	7.60	23.07
	Medical Textbooks	37.09	24.78	27.41	20.29	29.33	15.86	7.85	23.23
	Yahoo Answers	36.88	24.88	27.95	20.00	30.47	15.87	7.84	23.41
Qwen2.5-72B	w/o RAG	38.75	25.07	28.29	19.96	31.91	16.35	7.64	24.00
	BioASQ	40.53	25.08	27.95	19.95	32.08	16.35	7.65	24.23
	HealthCareMagic	40.12	25.22	27.84	19.28	32.85	16.37	7.42	24.16
	Medical Textbooks	39.26	25.23	28.13	20.60	31.78	16.32	7.63	24.14
	Yahoo Answers	40.47	25.21	28.48	19.55	32.93	16.44	7.76	24.41

Table 10: ROUGE-1 by model and retrieval corpus (open-ended datasets).

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	61.72	51.39	55.50	51.82	56.98	46.44	43.43	52.47
	BioASQ	62.88	52.44	55.04	50.84	56.88	47.07	44.78	52.85
	HealthCareMagic	57.43	52.56	54.25	49.98	56.72	47.17	43.52	51.66
	Medical Textbooks	59.87	52.40	54.81	50.53	56.03	47.05	44.45	52.16
	Yahoo Answers	58.37	52.29	55.13	49.41	56.61	47.03	43.98	51.83
LLaMA3.1-70B	w/o RAG	61.74	52.47	57.29	52.91	58.43	46.69	43.20	53.25
	BioASQ	63.96	52.40	54.45	52.27	56.84	46.63	44.30	52.97
	HealthCareMagic	60.41	52.58	55.53	53.46	58.35	46.55	44.36	53.03
	Medical Textbooks	61.08	52.44	54.78	53.15	56.45	46.53	44.34	52.68
	Yahoo Answers	60.87	52.31	55.02	52.54	57.38	46.57	44.37	52.72
Mistral-7B	w/o RAG	62.52	53.37	57.13	51.96	58.17	47.76	44.21	53.59
	BioASQ	63.06	53.49	55.84	51.23	57.33	48.08	45.52	53.51
	HealthCareMagic	61.77	54.04	56.87	51.41	58.29	48.10	45.20	53.67
	Medical Textbooks	61.05	53.43	56.09	52.10	57.10	48.05	45.20	53.29
	Yahoo Answers	61.76	53.47	56.77	51.22	58.11	48.20	45.32	53.55
Qwen2.5-7B	w/o RAG	61.28	52.02	55.81	50.94	56.07	46.07	42.91	52.16
	BioASQ	63.23	52.62	56.13	51.36	57.06	46.62	44.09	53.02
	HealthCareMagic	61.36	52.55	56.10	50.74	57.33	46.61	43.68	52.62
	Medical Textbooks	61.83	52.54	56.24	51.79	56.89	46.55	44.17	52.86
	Yahoo Answers	61.28	52.44	56.38	51.10	57.37	46.71	44.09	52.77
Qwen2.5-72B	w/o RAG	62.26	51.56	56.45	50.50	57.26	46.40	42.58	52.43
	BioASQ	64.30	51.92	56.60	50.98	57.86	46.63	43.34	53.09
	HealthCareMagic	63.05	52.07	56.40	50.26	58.04	46.69	42.94	52.78
	Medical Textbooks	63.07	51.93	56.81	51.54	57.75	46.57	43.29	52.99
	Yahoo Answers	63.56	51.88	56.73	50.53	58.20	46.72	43.33	52.99

Table 11: BERTScore by model and retrieval corpus (open-ended datasets).

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	29.84	16.50	21.84	16.05	19.09	14.75	11.96	18.58
	BioASQ	30.92	18.70	18.82	13.82	15.97	15.26	13.02	18.07
	HealthCareMagic	25.69	19.11	18.89	13.64	18.37	15.32	12.18	17.60
	Medical Textbooks	26.58	18.73	18.44	12.92	15.40	15.18	12.69	17.13
	Yahoo Answers	25.73	18.92	19.28	13.36	17.66	15.21	12.22	17.48
LLaMA3.1-70B	w/o RAG	30.51	18.85	20.65	16.62	19.02	14.93	11.89	18.92
	BioASQ	31.90	18.37	15.08	13.76	13.04	14.37	12.67	17.03
	HealthCareMagic	28.27	18.32	15.92	14.63	15.86	14.25	12.59	17.12
	Medical Textbooks	27.75	18.45	14.77	13.60	12.16	14.42	12.48	16.23
	Yahoo Answers	28.76	18.35	14.40	13.01	14.09	14.22	12.68	16.50
Mistral-7B	w/o RAG	29.90	19.29	21.74	15.24	18.92	15.79	12.88	19.11
	BioASQ	31.67	19.31	18.94	14.06	16.13	15.79	14.19	18.58
	HealthCareMagic	29.29	19.45	19.85	13.98	18.09	15.91	13.94	18.64
	Medical Textbooks	28.42	19.40	19.41	14.61	16.69	15.78	14.00	18.33
	Yahoo Answers	30.11	19.25	19.49	13.85	17.76	15.82	14.49	18.68
Qwen2.5-7B	w/o RAG	30.05	19.22	22.34	14.73	18.89	15.69	12.69	19.09
	BioASQ	31.30	19.35	20.66	14.56	18.15	15.74	13.05	18.97
	HealthCareMagic	29.74	19.28	20.91	14.10	19.02	15.68	12.91	18.81
	Medical Textbooks	30.02	19.40	20.46	14.58	17.75	15.77	13.28	18.75
	Yahoo Answers	29.19	19.09	20.49	14.09	19.21	15.63	13.38	18.73
Qwen2.5-72B	w/o RAG	31.11	19.12	22.27	14.45	20.09	15.98	12.60	19.37
	BioASQ	34.05	19.44	22.01	15.07	19.86	16.05	12.97	19.92
	HealthCareMagic	33.11	19.58	22.18	14.69	20.62	16.07	12.60	19.84
	Medical Textbooks	32.75	19.45	22.15	15.30	19.86	16.04	12.82	19.77
	Yahoo Answers	33.44	19.24	21.90	14.36	20.73	16.10	12.97	19.82

Table 12: METEOR by model and retrieval corpus (open-ended datasets).

Model	Retrieval Dataset	BioASQ	ChatDoctor/iCliniq	MashQA	MedicationQA	MedQuAD	MedRedQA	MeQSum	Average
LLaMA3.1-8B	w/o RAG	12.92	0.72	1.42	1.13	2.54	0.47	0.40	2.80
	BioASQ	19.08	0.77	1.52	1.00	1.99	0.48	0.53	3.62
	HealthCareMagic	10.52	0.82	1.39	0.99	2.34	0.48	0.50	2.43
	Medical Textbooks	12.99	0.78	1.49	1.02	1.83	0.48	0.49	2.73
	Yahoo Answers	10.79	0.76	1.53	0.97	2.16	0.47	0.44	2.45
LLaMA3.1-70B	w/o RAG	13.32	0.84	2.03	1.42	3.08	0.47	0.31	3.07
	BioASQ	19.76	0.85	1.56	1.19	1.73	0.49	0.39	3.71
	HealthCareMagic	14.16	0.92	1.72	1.57	2.58	0.48	0.44	3.12
	Medical Textbooks	15.92	0.87	1.47	1.53	1.67	0.48	0.41	3.19
	Yahoo Answers	13.97	0.92	1.54	1.43	2.03	0.49	0.47	2.98
Mistral-7B	w/o RAG	13.09	0.76	1.62	1.06	2.28	0.47	0.35	2.80
	BioASQ	17.25	0.83	1.67	1.07	2.12	0.51	0.45	3.41
	HealthCareMagic	13.74	0.96	1.69	1.12	2.35	0.53	0.43	2.97
	Medical Textbooks	12.71	0.81	1.58	1.15	2.01	0.50	0.45	2.74
	Yahoo Answers	14.32	0.86	1.75	1.14	2.44	0.52	0.46	3.07
Qwen2.5-7B	w/o RAG	11.83	0.68	1.65	1.03	1.99	0.43	0.33	2.56
	BioASQ	15.71	0.73	1.77	1.04	2.11	0.45	0.37	3.17
	HealthCareMagic	11.85	0.71	1.61	0.94	2.13	0.45	0.35	2.58
	Medical Textbooks	13.12	0.74	1.68	1.07	1.97	0.45	0.39	2.77
	Yahoo Answers	11.51	0.73	1.66	1.06	2.15	0.47	0.38	2.57
Qwen2.5-72B	w/o RAG	13.58	0.72	1.84	1.12	2.65	0.49	0.31	2.96
	BioASQ	16.43	0.76	1.83	1.10	2.59	0.51	0.34	3.37
	HealthCareMagic	14.51	0.78	1.80	0.99	2.88	0.52	0.32	3.11
	Medical Textbooks	14.67	0.77	1.89	1.12	2.62	0.50	0.34	3.13
	Yahoo Answers	14.79	0.75	1.92	1.04	2.95	0.53	0.36	3.19

Table 13: BLEU by model and retrieval corpus (open-ended datasets).

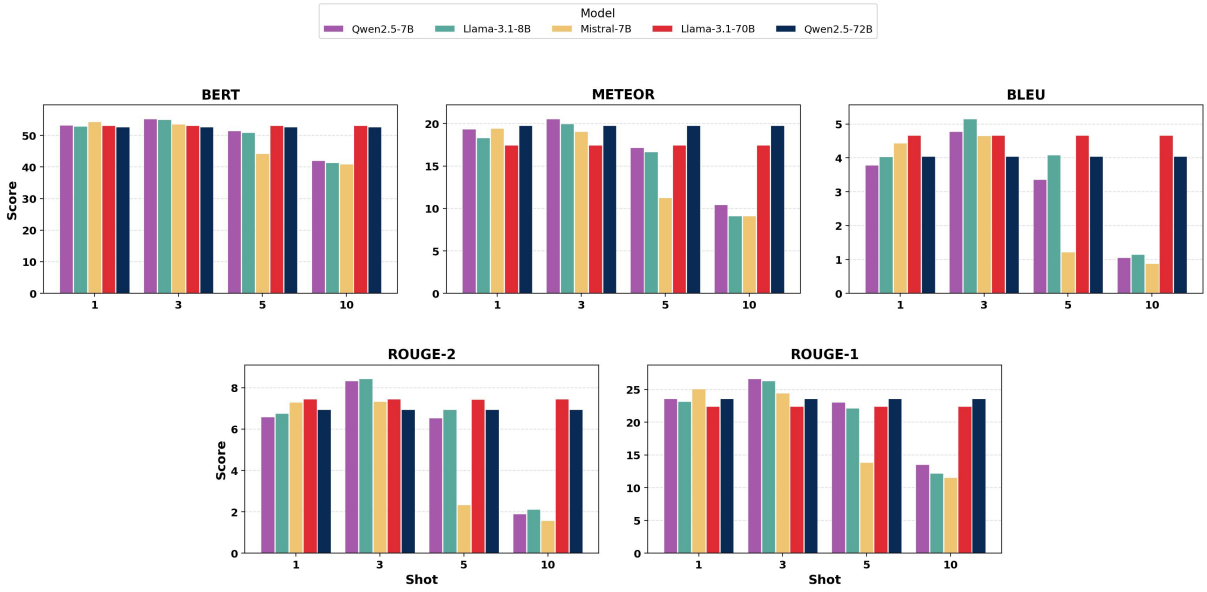


Figure 7: Additional open-ended metrics across shot counts (few-shot ablation).

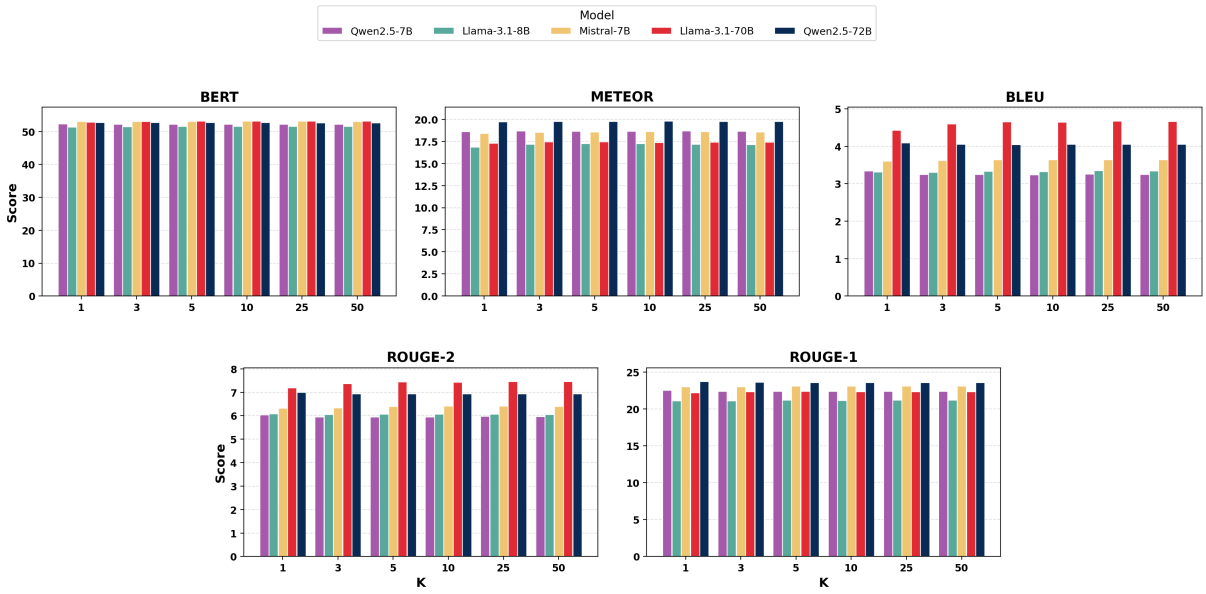


Figure 8: Additional open-ended metrics across top-k values.

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	22.20	22.63	22.32	21.75
	HealthCareMagic	21.09	21.70	21.48	21.48
	Medical Textbooks	21.48	21.31	21.75	21.40
	Yahoo Answers	21.30	21.66	21.43	20.55
	Average	21.52	21.83	21.74	21.30
LLaMA3.1-70B	BioASQ	21.58	22.40	21.82	21.68
	HealthCareMagic	22.50	22.63	22.90	22.62
	Medical Textbooks	21.30	21.54	21.81	21.56
	Yahoo Answers	21.64	21.98	22.25	21.56
	Average	21.76	22.14	22.20	21.85
Mistral-7B	BioASQ	23.60	23.88	23.59	23.31
	HealthCareMagic	23.98	23.97	24.00	24.07
	Medical Textbooks	23.31	23.35	23.20	23.32
	Yahoo Answers	23.88	23.90	24.06	24.01
	Average	23.69	23.77	23.71	23.68
Qwen2.5-7B	BioASQ	23.05	23.34	23.18	22.98
	HealthCareMagic	22.84	22.97	22.91	22.79
	Medical Textbooks	22.87	23.18	23.01	23.14
	Yahoo Answers	23.00	23.17	23.39	23.24
	Average	22.94	23.17	23.12	23.04
Qwen2.5-72B	BioASQ	24.03	24.09	23.91	24.01
	HealthCareMagic	23.83	23.91	23.90	24.13
	Medical Textbooks	24.03	24.05	23.90	23.97
	Yahoo Answers	24.00	24.11	24.37	24.22
	Average	23.97	24.04	24.02	24.08

Table 14: ROUGE-1 by retrieval method.

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	6.03	6.25	5.74	5.51
	HealthCareMagic	5.16	5.41	5.36	5.37
	Medical Textbooks	5.27	5.33	5.41	5.33
	Yahoo Answers	5.27	5.34	5.26	5.12
	Average	5.43	5.58	5.44	5.33
LLaMA3.1-70B	BioASQ	6.00	6.50	5.87	5.73
	HealthCareMagic	6.11	6.31	6.46	6.19
	Medical Textbooks	5.70	5.81	6.01	5.97
	Yahoo Answers	5.90	6.00	6.03	5.71
	Average	5.93	6.16	6.10	5.90
Mistral-7B	BioASQ	5.85	6.03	5.62	5.64
	HealthCareMagic	5.81	5.86	5.86	5.91
	Medical Textbooks	5.52	5.61	5.50	5.50
	Yahoo Answers	5.91	5.93	5.91	5.91
	Average	5.77	5.86	5.72	5.74
Qwen2.5-7B	BioASQ	5.48	5.77	5.52	5.36
	HealthCareMagic	5.25	5.36	5.35	5.36
	Medical Textbooks	5.34	5.59	5.45	5.45
	Yahoo Answers	5.47	5.52	5.47	5.48
	Average	5.39	5.56	5.45	5.41
Qwen2.5-72B	BioASQ	6.22	6.27	6.03	6.21
	HealthCareMagic	6.10	6.18	6.24	6.26
	Medical Textbooks	6.24	6.21	6.05	6.22
	Yahoo Answers	6.18	6.22	6.33	6.35
	Average	6.19	6.22	6.16	6.26

Table 15: ROUGE-2 by retrieval method.

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	3.32	3.59	3.10	2.81
	HealthCareMagic	2.11	2.36	2.38	2.31
	Medical Textbooks	2.47	2.61	2.70	2.48
	Yahoo Answers	2.42	2.43	2.34	2.20
	Average	2.58	2.75	2.63	2.45
LLaMA3.1-70B	BioASQ	3.26	3.67	3.15	2.94
	HealthCareMagic	2.90	2.89	3.12	2.94
	Medical Textbooks	2.81	2.80	2.85	3.10
	Yahoo Answers	2.73	2.86	2.96	2.77
	Average	2.93	3.06	3.02	2.94
Mistral-7B	BioASQ	3.19	3.38	2.93	2.68
	HealthCareMagic	2.77	2.89	2.92	2.94
	Medical Textbooks	2.66	2.70	2.69	2.56
	Yahoo Answers	2.95	3.01	2.95	2.98
	Average	2.89	3.00	2.87	2.79
Qwen2.5-7B	BioASQ	2.76	3.14	2.67	2.46
	HealthCareMagic	2.39	2.45	2.54	2.50
	Medical Textbooks	2.53	2.76	2.56	2.46
	Yahoo Answers	2.52	2.54	2.51	2.55
	Average	2.55	2.72	2.57	2.49
Qwen2.5-72B	BioASQ	3.24	3.34	3.05	3.06
	HealthCareMagic	2.95	2.95	3.10	3.10
	Medical Textbooks	3.12	3.01	2.93	3.01
	Yahoo Answers	3.01	3.11	3.14	3.18
	Average	3.08	3.10	3.06	3.09

Table 16: BLEU by retrieval method.

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	17.19	17.54	17.55	17.27
	HealthCareMagic	16.53	17.02	17.17	17.05
	Medical Textbooks	16.77	16.76	17.06	16.96
	Yahoo Answers	17.04	17.32	17.27	16.62
	Average	16.88	17.16	17.26	16.98
LLaMA3.1-70B	BioASQ	16.11	16.55	16.32	16.17
	HealthCareMagic	16.56	16.70	17.07	16.82
	Medical Textbooks	15.65	15.77	16.17	15.86
	Yahoo Answers	15.94	16.04	16.36	16.08
	Average	16.06	16.26	16.48	16.23
Mistral-7B	BioASQ	18.03	18.38	18.13	18.00
	HealthCareMagic	18.33	18.36	18.44	18.45
	Medical Textbooks	18.12	18.08	18.14	18.12
	Yahoo Answers	18.34	18.39	18.31	18.55
	Average	18.20	18.30	18.25	18.28
Qwen2.5-7B	BioASQ	18.37	18.74	18.59	18.43
	HealthCareMagic	18.48	18.44	18.70	18.61
	Medical Textbooks	18.40	18.67	18.49	18.44
	Yahoo Answers	18.53	18.46	18.64	18.59
	Average	18.44	18.58	18.61	18.52
Qwen2.5-72B	BioASQ	19.63	19.76	19.56	19.52
	HealthCareMagic	19.55	19.65	19.81	19.57
	Medical Textbooks	19.65	19.61	19.44	19.61
	Yahoo Answers	19.52	19.53	19.65	19.71
	Average	19.59	19.64	19.61	19.60

Table 17: METEOR by retrieval method.

Model	Retrieval Dataset	BM25	Hybrid	MedCPT	TF-IDF
LLaMA3.1-8B	BioASQ	52.48	52.72	52.54	52.08
	HealthCareMagic	50.92	51.33	51.24	51.37
	Medical Textbooks	51.75	51.82	52.14	51.73
	Yahoo Answers	51.43	51.71	51.53	50.80
	Average	51.65	51.89	51.86	51.49
LLaMA3.1-70B	BioASQ	52.36	52.88	52.52	52.07
	HealthCareMagic	52.69	52.84	52.94	52.78
	Medical Textbooks	52.02	52.38	52.62	52.25
	Yahoo Answers	52.25	52.62	52.65	52.03
	Average	52.33	52.68	52.68	52.28
Mistral-7B	BioASQ	53.25	53.39	53.24	53.04
	HealthCareMagic	53.47	53.55	53.53	53.56
	Medical Textbooks	53.12	53.16	53.04	53.03
	Yahoo Answers	53.40	53.43	53.50	53.39
	Average	53.31	53.39	53.33	53.26
Qwen2.5-7B	BioASQ	52.72	52.96	52.80	52.53
	HealthCareMagic	52.37	52.45	52.51	52.37
	Medical Textbooks	52.52	52.81	52.73	52.64
	Yahoo Answers	52.65	52.64	52.75	52.60
	Average	52.57	52.71	52.70	52.54
Qwen2.5-72B	BioASQ	52.93	53.01	52.87	52.90
	HealthCareMagic	52.60	52.70	52.74	52.69
	Medical Textbooks	52.89	52.87	52.84	52.85
	Yahoo Answers	52.74	52.85	52.92	52.86
	Average	52.79	52.86	52.84	52.83

Table 18: BERTScore by retrieval method.

CROSSDDI: Cross-Source Evidence-Grounded Drug–Drug Interaction Verification

Hui Wang **Bohao Chu** **Norbert Fuhr**
University Duisburg-Essen
{hui.wang.0111@stud, bohao.chu, norbert.fuhr}@uni-due.de

Abstract

LLM-based drug–drug interaction (DDI) assessment remains difficult to audit when predictions are not explicitly tied to evidence. While retrieval-augmented generation (RAG) improves grounding, predictions are not guaranteed to be entailed by retrieved items. We present CROSSDDI, a verification-first framework that separates LLM-based evidence extraction from deterministic arbitration over DrugBank and PubMed, requiring positive predictions to be linked to explicit supporting evidence. Evaluated on 1,000 DDInter 2.0 pairs under a positive–unlabeled setting, CROSSDDI achieves recall of 0.576–0.593 over confirmed positives with interaction prediction rates comparable to RAG, while reducing cross-backbone variation (0.018 vs. 0.066). Analysis identifies literature evidence acquisition and attribution as the primary bottleneck: PubMed retrieval covers only 40.5% of confirmed positives, and Path B-only evidence is substantially less reliable than structured evidence. These results suggest that verification-first architectures can improve traceability and backbone consistency, while broader and more reliable literature evidence is needed to extend coverage beyond structured sources.

1 Introduction

Users increasingly turn to conversational AI for medication-related questions, including safety-critical queries such as "Can I take warfarin with aspirin?" (White et al., 2025). Yet large language models (LLMs) can produce hallucinated medical claims (Singhal et al., 2023), and in drug–drug interaction (DDI) assessment, missed or incorrect identification of interactions may lead to adverse events (Huang et al., 2025). Reliable DDI assessment therefore requires every prediction to be traceable to explicit evidence.

Retrieval-augmented generation (RAG) augments language models with retrieved text passages from external sources (Asai et al., 2024) but does not guarantee that generated predictions are faithfully grounded in the retrieved evidence (Wu et al., 2024). Neural DDI models (He et al., 2022; Wang et al., 2025a) predict interactions from molecular features, but operate on structured inputs and do not directly support verification of free-form claims against source-level evidence.

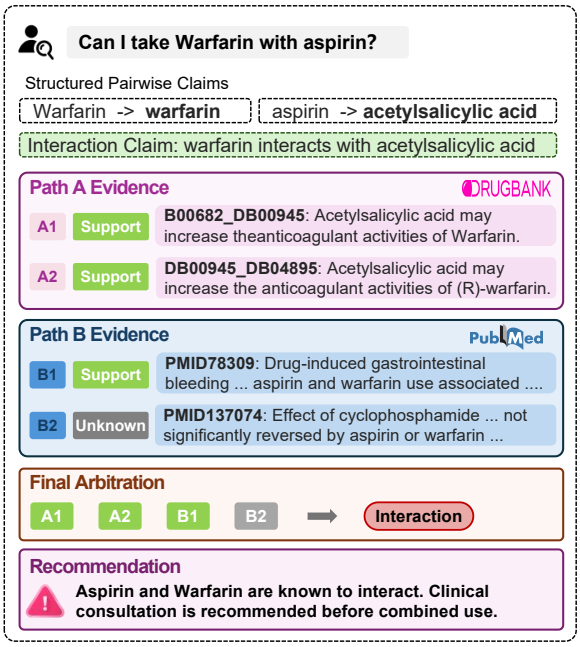


Figure 1: End-to-end example of CROSSDDI for the query "Can I take Warfarin with aspirin?" DrugBank provides structured evidence and PubMed complementary evidence; deterministic arbitration yields an *Interaction* verdict linked to explicit evidence.

We introduce DDI assessment as **evidence-grounded claim verification** and propose CROSSDDI, which separates LLM-based evidence extraction from deterministic decision-making over two complementary sources: DrugBank (Knox et al., 2024) for broad structured coverage and PubMed for long-tail or recent cases. LLMs assign a stance from each source; a rule-based arbitration layer combines these stances to produce a final ver-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 911–919
July 3-4, 2026 ©2026 Association for Computational Linguistics

dict, *guaranteeing* that every positive prediction is backed by explicit evidence (Figure 1).

Curated DDI resources such as DDInter (Tian et al., 2025) primarily provide annotated interaction records, while non-interactions are typically unlabeled. In such settings, drug–drug interaction modeling is commonly formulated as a positive–unlabeled (PU) learning problem (Bi et al., 2018). Across three LLM backbones (Qwen2.5-7B, GPT-4o, and Qwen3-235B-A22B (Qwen et al., 2025; Hurst et al., 2024; Yang et al., 2025)), CROSSDDI achieves recall (0.576–0.593) comparable to generation-based RAG baselines while ensuring full evidence traceability, and reduces cross-backbone variation (recall spread 0.018 vs. 0.066). Our analysis traces the remaining recall gap to sparse literature-side evidence acquisition.

Contributions. (i) We formulate DDI assessment as evidence-grounded claim verification over drug pairs under the positive–unlabeled (PU) setting, requiring explicit evidence support. (ii) We introduce CROSSDDI, a verification-first framework that separates LLM-based evidence extraction from deterministic decision-making, enabling explicit grounding, traceability, and consistent predictions across LLM backbones. (iii) Through staged error analysis, we identify literature-side evidence acquisition as a primary bottleneck for recall.

2 Related Work

DDI prediction and extraction. Prior DDI work includes deep learning methods (e.g., SSF-DDI (Zhu et al., 2024), CNN-DDI (Zhang et al., 2022)) and knowledge-graph-based approaches (e.g., LLM-DDI (Li et al., 2026), KGDB-DDI (Zhao et al., 2025)), which learn drug representations from molecular, structural, or relational data for interaction prediction. These methods focus on prediction rather than verification and do not provide evidence-grounded justification for individual claims. Recent LLM-based approaches have been applied to medical QA settings (Wang et al., 2025b), but remain prone to hallucinated or unsupported claims (Wang et al., 2025b; Sicard et al., 2025).

Biomedical verification. Prior work on scientific and medical fact verification, including SciFact, MultiVerS, HealthFC, and BioFactCheck (Wadden et al., 2022a,b; Vladika et al., 2024; Lamichhane et al., 2023), assesses whether claims are supported or refuted by retrieved textual evidence. These approaches typically operate over abstract- or document-level evidence and treat claims independently, without modeling structured pairwise interactions or integrating evidence across curated knowledge and clinical literature. CROSSDDI instead targets structured pairwise drug interaction claims and combines evidence from curated databases and clinical literature through verification with deterministic arbitration.

Biomedical RAG. Biomedical RAG systems (e.g., Almanac (Zakka et al., 2024)) improve grounding by incorporating retrieved evidence, but typically rely on the generator to reconcile evidence during generation. As a result, predictions are not guaranteed to be entailed by specific retrieved items (Wu et al., 2024). In contrast, CROSSDDI separates evidence extraction from deterministic arbitration, enabling traceable decisions under heterogeneous and incomplete evidence.

3 CrossDDIPipeline

CROSSDDI proceeds in four stages: (1) entity resolution and claim construction, (2) DrugBank verification (Path A), (3) PubMed verification (Path B), and (4) deterministic arbitration (see Appendix A). LLMs handle only evidence extraction and attribution; final verdicts follow a fixed rule.

3.1 Entity Resolution and Claim Construction

Given a query q, drug mentions are normalized via case folding, suffix stripping, and alias matching against DrugBank synonyms, yielding a canonical pair $\phi(q) = (d_1, d_2),\ d_1 \leq_{\text{lex}} d_2$. We then construct the binary claim $c = \text{Interacts}(d_1, d_2)$.

This normalization reduces surface variation but may fail for complex multi-token drug names; such cases are treated as part of the overall pipeline error profile analyzed later.

3.2 Path A: Structured DrugBank Verification

Path A queries a local DrugBank (Knox et al., 2024) index with a ranked retrieval procedure. The primary channel is an exact symmetric pair-key match over normalized DrugBank identifiers. Auxiliary lexical channels retrieve lower-specificity records whose descriptions mention both drug names or whose counterpart field matches one queried drug. Records are ranked by specificity, and the top K ($K = 3$) are retained.

Let $R_A^{(K)}$ denote the ranked top-K DrugBank output, and let $\text{Exact}(r, d_1, d_2)$ indicate that record

r is an exact pair-key match for the normalized pair (d_1, d_2). The path-level stance is defined as

$$y_A = \begin{cases} \textit{Support} & \exists r \in R_A^{(K)} : \mathrm{Exact}(r, d_1, d_2), \\ \textit{Unknown} & \text{otherwise.} \end{cases}$$

Exact pair matches are treated as structured support evidence, whereas lexical matches provide auxiliary context only. The absence of an exact match is interpreted as unresolved rather than negative evidence. Path A does not emit *Contradict*, consistent with the PU setting.

3.3 Path B: Literature-grounded Verification

Path B performs sentence-level verification over PubMed. Synonym-expanded queries retrieve up to three articles, from which an LLM extracts at most six claim-relevant sentences and assigns stance labels in $\{\textit{Support}, \textit{Contradict}, \textit{Unknown}\}$. Post-extraction filtering reduces spurious signals: sentences that merely co-mention the drugs are downgraded to *Unknown*, and *Contradict* is kept only when both explicit negation and interaction cues are present. These heuristics prioritize precision over recall. Let E_B^+ and E_B^- denote the filtered *Support* and *Contradict* sets. The path-level stance is defined as

$$y_B = \begin{cases} \textit{Support} & E_B^+ \neq \emptyset \wedge E_B^- = \emptyset, \\ \textit{Contradict} & E_B^- \neq \emptyset \wedge E_B^+ = \emptyset, \\ \textit{Unknown} & \text{otherwise.} \end{cases}$$

In practice, retrieval coverage is limited and many queries default to *Unknown*, making Path B a sparse but complementary signal. Despite this, it provides a literature channel that links decisions to explicit textual statements, supporting interpretability and enabling inspection of model outputs. Full prompt details are provided in Appendix F.

3.4 Deterministic Arbitration

Path-level outputs (y_A, y_B) are combined via a fixed rule:

$$y = \begin{cases} \text{Interaction} & y_A = \mathrm{Sup} \vee y_B = \mathrm{Sup}, \\ \text{Non-Interaction} & y_A = \mathrm{Unk} \wedge y_B = \mathrm{Con}, \\ \text{Unknown} & \text{otherwise.} \end{cases}$$

The rule is asymmetric: any supporting signal yields *Interaction*, reflecting the higher cost of missing harmful interactions, while *Non-Interaction* requires explicit contradiction without support.

A fixed rule is preferred over learned fusion given the absence of negative supervision in the PU setting. Positive predictions are linked by design to retrieved supporting evidence, though their correctness depends on extraction quality.

4 Experiments

4.1 Setup

Dataset and PU setting. We evaluate on 1,000 DDInter 2.0 pairs (Tian et al., 2025) (seed 42): 797 confirmed interactions and 203 unlabeled pairs. Because DDInter provides no verified *Non-Interaction* labels, evaluation is positive–unlabeled (PU) (Bi et al., 2018); precision and false-positive rate are not identifiable.

Metrics. We use **recall over the 797 confirmed positives** as the primary metric and report interaction prediction rate (P.Int) over all pairs, with P.Unk and P.NI describing abstention behavior. Backbone variation is summarized by recall spread. Bootstrap 95% confidence intervals (2,000 resamples) are reported for P.Int, not recall. Accuracy is reported in Appendix D only for completeness.

Pseudo-negative analysis. We form pseudo-negative (PN) subsets from unlabeled pairs with both drugs resolved and no Path A DrugBank match ($n = 91$–93 per suite). Under a closed-world assumption, *Interaction* predictions on these pairs approximate false positives, so PN-FPR is used only for within-suite comparison. Wilson 95% CIs reflect uncertainty within the PN subset, not clinical false-positive uncertainty.

Baselines. We compare CROSSDDI with *LLM-only*, *RAG*, *Path A*, and *Path B* across Qwen2.5-7B, GPT-4o, and Qwen3-235B. Main comparisons use normalized pair inputs after claim construction, so all retrieval-based conditions use the same DrugBank and PubMed evidence retrieved for CROSSDDI. RAG receives this evidence and generates a final verdict using a fixed prompt (Appendix G). These comparisons are matched at the evidence level, isolating differences in reasoning and arbitration; full parsing and claim-construction failures are analyzed separately. Because Path A is deterministic, its outputs are identical across backbones.

4.2 Results

Overall performance. CROSSDDI achieves the highest recall on two of three backbones (0.580, 0.593) and is comparable on the third (0.576 vs.

Suite	PN n	LLM	RAG	Path A	Path B	CrossDDI
Qwen2.5	91	27.5 [19.4, 37.4]	3.3 [1.1, 9.2]	0.0 [0.0, 4.1]	11.0 [6.1, 19.1]	12.1 [6.9, 20.4]
GPT-4o	93	6.5 [3.0, 13.4]	6.5 [3.0, 13.4]	0.0 [0.0, 4.0]	14.0 [8.4, 22.5]	19.4 [12.6, 28.5]
Qwen3	92	5.4 [2.3, 12.1]	13.0 [7.6, 21.4]	0.0 [0.0, 4.0]	14.1 [8.4, 22.7]	17.4 [11.0, 26.4]

Table 1: Pseudo-negative FPR (heuristic). Percentages with Wilson 95% CIs in brackets.

Model	Setting	P.Int	P.Int 95% CI	Recall	P.Unk	P.NI
Qwen2.5-7B	LLM-only	0.385	[.355,.413]	0.386	0.510	0.104
	RAG	0.546	[.513,.578]	0.566	0.434	0.019
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.051	[.037,.067]	0.041	0.937	0.004
	CROSSDDI	**0.556**	[.524,.587]	**0.580**	0.435	0.000
GPT-4o	LLM-only	0.351	[.320,.381]	0.407	0.311	0.338
	RAG	0.478	[.445,.511]	0.512	0.517	0.003
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.089	[.070,.110]	0.077	0.868	0.001
	CROSSDDI	**0.575**	[.543,.606]	**0.593**	0.424	0.000
Qwen3-235B	LLM-only	0.320	[.291,.350]	0.384	0.580	0.100
	RAG	0.564	[.531,.596]	**0.578**	0.432	0.004
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.083	[.064,.104]	0.074	0.908	0.001
	CROSSDDI	0.557	[.525,.588]	0.576	0.427	0.001

Table 2: Interaction prediction rate (P.Int) over all pairs, bootstrap confidence intervals for P.Int, and recall over confirmed positives. P.Unk and P.NI denote prediction rates for *Unknown* and *Non-Interaction*.

0.578 for RAG) under matched retrieval inputs. Interaction prediction rates are similar (0.556–0.575 vs. 0.478–0.564), indicating that gains are not driven by substantially more frequent positive predictions. Path A alone already provides strong coverage (recall = 0.565, Table 2). Among the 1,000 evaluation pairs, 558 (55.80%) have direct DrugBank pair coverage under the Path A resolver. Recall on confirmed positive pairs with direct DrugBank pair coverage reaches 85.81%, compared with 18.98% on confirmed positive pairs without such coverage, indicating that structured database coverage strongly influences end-to-end performance (Appendix E).

Backbone consistency. CROSSDDI exhibits lower cross-backbone variation (recall spread 0.018) than RAG (0.066), indicating more consistent behavior across models. This difference is partly attributable to the dominance of structured evidence in Path A, which reduces the influence of LLM-dependent variation from Path B during decision making.

Literature evidence. Path B alone yields low recall (0.041–0.077) and predicts *Unknown* for most pairs. PubMed retrieval applies to only 40.5% of confirmed positives, making literature-side evidence acquisition the primary bottleneck. Additional loss arises from entity resolution, query construction, and extraction reliability. Despite these limitations, Path B provides complementary evidence in cases where structured lookup abstains, linking decisions to explicit textual statements beyond curated sources. At the system level, the support-favoring arbitration rule leads to higher PN-FPR than RAG across all pseudo-negative suites (Table 1), reflecting an inherent recall–FPR tradeoff in the current design.

Evidence quality. Manual evaluation of 100 stratified *Interaction* predictions shows overall stance correctness of 84.0%, with 85.0% strict evidence relevance and 95.0% lenient evidence relevance. Evidence quality varies substantially by trigger source: Path A-only and both-path cases show high stance correctness (96.0% and 100.0%, respectively), while Path B-only cases are weaker (44.0% correctness). The main failure mode for Path B-only cases is selection of topically related but non-interaction statements during retrieval and extraction. This evaluation uses a single annotator and should be treated as preliminary; future work should include multi-annotator evaluation and agreement analysis to assess annotation reliability.

5 Conclusion

CROSSDDI shows that separating evidence extraction from deterministic decision-making yields more consistent behavior across LLM backbones (recall spread 0.018 vs. 0.066 for RAG). Under matched retrieval, CROSSDDI achieves higher recall on two of three backbones and comparable recall on the third, with interaction prediction rates similar to RAG. Coverage analysis shows that end-to-end recall is strongly influenced by direct DrugBank coverage: recall over confirmed positives reaches 85.81% on covered pairs but drops to 18.98% on uncovered pairs. The literature path remains limited by retrieval coverage (40.5% of confirmed positives) and extraction reliability. These results suggest that verification-first architectures can improve traceability and backbone consistency, while better literature retrieval and evidence quality are needed for coverage beyond structured sources.

Limitations

Positive–unlabeled evaluation. DDInter provides confirmed positive pairs but no verified *Non-Interaction* labels, resulting in a PU setting in which precision and false-positive rate are not identifiable. Our pseudo-negative FPR is therefore heuristic and intended only for relative comparison within each backbone suite.

Curated-resource overlap. Because DDInter and DrugBank are both curated DDI resources, their interaction coverage may partially overlap. Quantitative analysis suggests a strong dependence on direct DrugBank coverage: recall drops from 85.81% to 18.98% when moving from pairs with direct DrugBank coverage to those without, on the same 1,000-pair evaluation set used in the reported experiments (Appendix E, Table 4). The strong Path A results should therefore be interpreted primarily as evidence of structured database coverage, and the extent to which current results reflect independent cross-source verification should be interpreted cautiously.

Incomplete and imbalanced evidence. Literature retrieval covers only 40.5% of confirmed positives, and a much smaller fraction yields *Support* evidence, reflecting limited coverage and attribution challenges. In contrast, structured knowledge (DrugBank) provides the dominant signal for end-to-end performance. As a result, *Unknown* predictions may reflect missing evidence rather than true uncertainty, and current behavior is largely driven by structured sources.

LLM dependence and task scope. Although final arbitration is deterministic, Path B relies on LLMs for sentence selection and stance attribution, introducing model-dependent variability at the extraction stage. Furthermore, CROSSDDI formulates DDI assessment as pairwise binary verification and does not account for dosage, route, or patient-specific factors.

Future Work

Evaluation and benchmarks. A key limitation of current DDI resources is the absence of verified non-interaction labels. Constructing benchmarks with both confirmed interactions and validated negatives would enable reliable evaluation of precision and false-positive control in evidence-grounded settings.

Retrieval and evidence coverage. Improving literature-side coverage remains critical. Future work includes expanding retrieval to broader sources (e.g., full-text articles and regulatory data) and improving drug-name normalization to reduce loss from entity resolution and retrieval failures.

Stronger baselines and comparisons. Further evaluation against stronger RAG variants (e.g., few-shot or chain-of-thought prompting) and neural DDI models under matched retrieval conditions is needed to better characterize the strengths and limitations of the verification-first design.

Ethics Statement

CROSSDDI is a research prototype for evidence-grounded DDI assessment and is not intended to replace professional clinical judgment. DrugBank data were accessed under an academic research license. No patient-level or personally identifiable data were used in this study.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xin Bi, He Ma, Jianhua Li, Yuliang Ma, and Deyang Chen. 2018. A positive and unlabeled learning framework based on extreme learning machine for drug-drug interactions discovery. *Journal of Ambient Intelligence and Humanized Computing*.

Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. 2022. 3dgt-ddi: 3d graph and text based neural network for drug–drug interaction prediction. *Briefings in Bioinformatics*, 23(3):bbac134.

Wenzhun Huang, Xiao Wang, Yunhao Chen, Changqing Yu, and Shanwen Zhang. 2025. Advancing drug-drug interactions research: integrating ai-powered prediction, vulnerable populations, and regulatory insights. *Frontiers in Pharmacology*, 16:1618701.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Craig Knox, Michael Wilson, Christian M. Klinger, et al. 2024. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275.

Prajwol Lamichhane, Indika Kahanda, Xudong Liu, Karthikeyan Umapathy, Sandeep Reddivari, Catherine Christie, Andrea Arikawa, and Jenifer Ross. 2023. Poster: Biofactcheck: Exploring the feasibility of explainable automated inconsistency detection in biomedical and health literature. In *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 196–197.

Dongxu Li, Yue Yang, Ziwen Cui, Hengchuang Yin, Pengwei Hu, and Lun Hu. 2026. Llm-ddi: Leveraging large language models for drug-drug interaction prediction on biomedical knowledge graph. *IEEE Journal of Biomedical and Health Informatics*, 30(1):773–781.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Justine Sicard, François Montastruc, Coline Achalme, Annie Pierre Jonville-Bera, Paul Songue, Marina Babin, Thomas Soeiro, Pauline Schiro, Claire de Canecaude, and Romain Barus. 2025. Can large language models detect drug–drug interactions leading to adverse drug reactions? *Therapeutic Advances in Drug Safety*, 16:20420986251339358.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yao Tian, Jiacai Yi, Ningning Wang, Chengkun Wu, Jinfu Peng, Shao Liu, Guoping Yang, and Dongsheng Cao. 2025. Ddinter 2.0: an enhanced drug interaction resource with expanded data coverage, new interaction types, and improved user interface. *Nucleic Acids Research*, 53(D1):D1356–D1362.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

GuiShen Wang, Hui Feng, and Chen Cao. 2025a. Birnn-ddi: A drug-drug interaction event type prediction model based on bidirectional recurrent neural network and graph2seq representation. *Journal of Computational Biology*, 32(2):198–211.

Sheng Wang, Fangyuan Zhao, Dechao Bu, Yunwei Lu, Ming Gong, Hongjie Liu, Zhaohui Yang, Xiaoxi Zeng, Zhiyuan Yuan, Baoping Wan, et al. 2025b. Lins: A general medical q&a framework for enhancing the quality and credibility of llm-generated responses. *Nature communications*, 16(1):9076.

Christopher A White, Yehuda A Masturov, Eric Haunschild, Evan Michaelson, Dave R Shukla, and Paul J Cagle. 2025. Can chatgpt reliably answer the most common patient questions regarding total shoulder arthroplasty? *Journal of Shoulder and Elbow Surgery*, 34(5):e254–e264.

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9390–9406, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Chengcheng Zhang, Yao Lu, and Tianyi Zang. 2022. Cnn-ddi: a learning-based method for predicting drug–drug interactions using convolutional neural networks. *BMC Bioinformatics*, 23(Suppl 1):88.

Changpeng Zhao, Dongfang Han, Zicheng Zuo, and Turdi Tohti. 2025. Kgdb-ddi: Knowledge graph-based drug background data fusion model for drug–drug interaction prediction. *Artificial Intelligence in Medicine*, 168:103225.

Jing Zhu, Chao Che, Hao Jiang, Jian Xu, Jiajun Yin, and Zhaoqian Zhong. 2024. Ssf-ddi: a deep learning method utilizing drug sequence and substructure features for drug–drug interaction prediction. *BMC bioinformatics*, 25(1):39.

A Pipeline Overview

B Implementation Details for Path A

Retrieval parameter. The parameter K bounds the number of DrugBank records returned per query. We set $K = 3$ by default, yielding a compact evidence set. Records are ranked by specificity: exact pair-key matches are ranked before lexical description matches and lower-specificity counterpart matches. Let $R_A^{(K)}$ denote this retained ranked output. Path A emits *Support* iff an exact pair-key match appears in $R_A^{(K)}$; otherwise it emits *Unknown*. Thus K controls both the retained evidence set and the retrieval output used for Path A stance assignment.

Interpretive reporting module. The retained DrugBank records are additionally processed by an LLM-based summarization module that produces structured summaries and natural-language descriptions. These outputs are used for interpretation only and do not affect the deterministic Path A stance defined in §3.2.

C Implementation Details for Path B

Alias expansion. Path B expands normalized drug names using a small fixed alias map covering common generic and brand-name variants. When no aliases are available, retrieval uses the normalized name.

Retrieval scope. Path B issues synonym-expanded queries, aggregates PubMed results, deduplicates by PMID, and retains the first three unique articles D_B.

Joint evidence extraction. The retrieved documents are processed by an LLM-based module (Prompt B.1) that extracts up to six candidate sentences, each labeled as *Support*, *Contradict*, or *Unknown*, together with a short rationale.

Post-extraction guardrails. Two deterministic rules are applied after extraction. (i) *Pairwise relevance*: sentences that do not match both queried drugs are downgraded to *Unknown*. (ii) *Contradiction filter*: *Contradict* is retained only when explicit negation and interaction cues are both present; otherwise it is downgraded to *Unknown*. These rules favor precision over recall.

Guardrail cue lists. The following cue lists are fixed across all experiments. Matching is case-insensitive and implemented as substring matching over the extracted sentence.

Explicit pairwise negation cues (used to retain *Contradict* labels): `no interaction`, `no significant interaction`, `no clinically meaningful interaction`, `did not identify an interaction`, `does not interact`, `no effect on`, `not associated with an interaction`.

Interaction-related cues (required alongside a negation cue to retain *Contradict*): `interaction`, `interact`, `coadministration`, `co-administration`, `concomitant`, `combined`, `bleeding risk`, `toxicity`, `exposure`, `serum level`, `metabolism`, `excretion`, `anticoagulant`, `inr`, `monitor`, `contraind`.

Interpretive reporting module. Retrieved documents are also summarized by an LLM for interpretive purposes. These summaries do not affect path-level stance or final decisions.

D Accuracy Under PU Evaluation

Table 3 reports accuracy for completeness. Accuracy is not a reliable measure under PU evaluation, as unlabeled pairs may contain true interactions and are treated as negatives.

E Structured Coverage Analysis

DDInter and DrugBank are both curated DDI resources, and their interaction coverage may partially overlap, potentially contributing to strong performance for systems relying on structured DrugBank evidence. We therefore analyze direct DrugBank coverage in the randomly sampled 1,000-pair DDInter evaluation set.

Under the deterministic resolver used by Path A, 558 pairs (55.80%) have a direct DrugBank pair match and 442 pairs (44.20%) do not. We then evaluate recall over confirmed positives separately

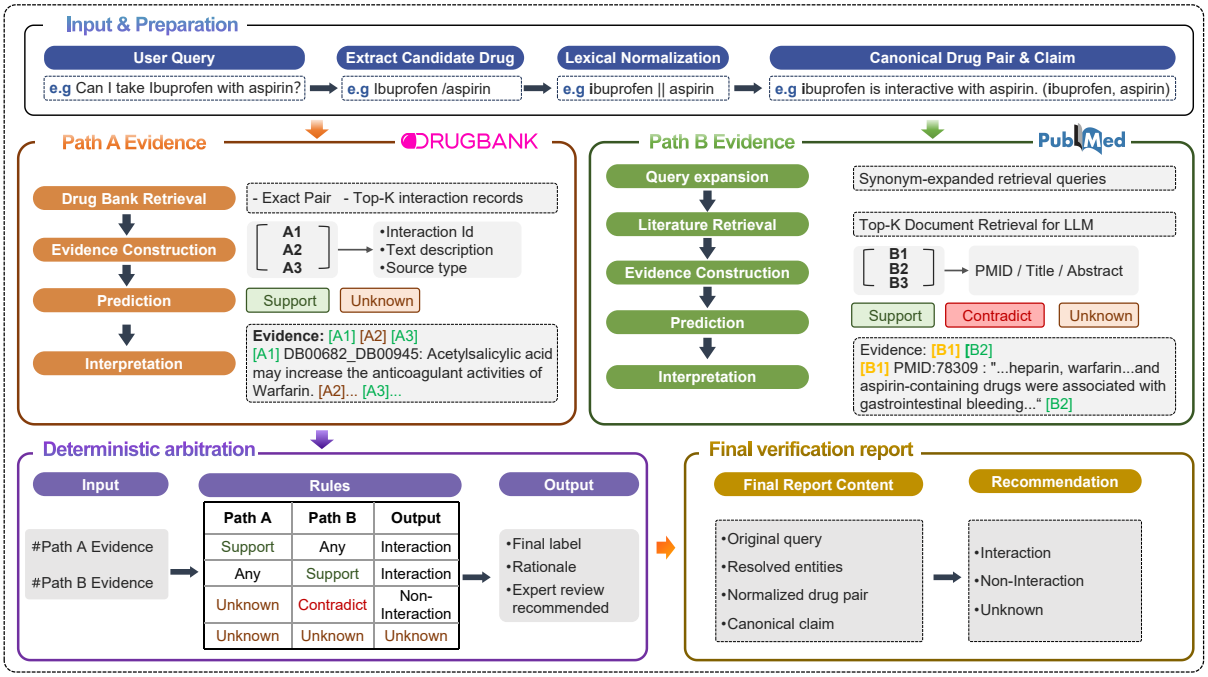


Figure 2: CROSSDDI pipeline. A drug-pair query is normalized into a canonical interaction claim and verified along two evidence paths: DrugBank (Path A) and PubMed (Path B), which provide structured and literature-based evidence, respectively. A deterministic arbitration rule combines path-level outputs into a final verdict linked to explicit supporting evidence.

for the two subsets. With direct DrugBank coverage, CROSSDDI recalls 399 of 465 positives (85.81%); without such coverage, it recalls 63 of 332 positives (18.98%). This gap suggests that structured database coverage strongly influences observed end-to-end recall.

F Prompt Template for Path B Evidence Extraction

Prompt B.1 is the only LLM component whose outputs directly influence the final decision, as it determines the evidence sets used for stance assignment.

Prompt B.1: Literature Joint Extraction.

```
You are an expert clinical pharmacologist
    performing joint
evidence extraction and stance labeling for
    PubMed drug-drug
interaction verification.
From the provided Top-K PubMed documents,
    extract the most
claim-relevant evidence sentences and assign
    each one a label:
Support, Contradict, or Unknown.

Rules:
- Prefer sentences that explicitly describe
    interaction,
  coadministration risk, bleeding risk, toxicity,
      exposure
  change, monitoring concern, or lack of
      interaction.
- Do not select sentences that merely co-mention
    both drugs
  unless the sentence discusses their pairwise
      relationship.
- Use Contradict only when the sentence
    explicitly denies a
  pairwise interaction or pairwise effect.
- Use Unknown when the sentence is indirect,
    weak,
  not pair-specific, or insufficient to resolve
      the claim.
- Return at most the requested number of
    sentences.
- Return valid JSON only.
- Use exactly this schema:
  {"selected_evidence":[{"pmid":"...","title
      ":"...",
  "sentence":"...","score":0.0,
  "source_type":"abstract_sentence",
  "label":"Support|Contradict|Unknown",
  "rationale":"one concise evidence-grounded
      sentence",
  "selection_rationale":"why this sentence was
      selected"}]}.
- Do not wrap the JSON in markdown fences.

Canonical claim: {claim}
Maximum sentences: {max_sentences}

Top-K PubMed documents:
{context}
```

Model	Setting	P.Int	P.Int 95% CI	Recall	P.Unk	P.NI	Acc.
Qwen2.5-7B	LLM-only	0.385	[.355,.413]	0.386	0.510	0.104	0.404
	RAG	0.546	[.513,.578]	0.566	0.434	0.019	0.553
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.051	[.037,.067]	0.041	0.937	0.004	0.215
	CROSSDDI	**0.556**	[.524,.587]	**0.580**	0.435	0.000	0.568
GPT-4o	LLM-only	0.351	[.320,.381]	0.407	0.311	0.338	0.372
	RAG	0.478	[.445,.511]	0.512	0.517	0.003	0.537
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.089	[.070,.110]	0.077	0.868	0.001	0.225
	CROSSDDI	**0.575**	[.543,.606]	**0.593**	0.424	0.000	0.574
Qwen3-235B	LLM-only	0.320	[.291,.350]	0.384	0.580	0.100	0.465
	RAG	0.564	[.531,.596]	**0.578**	0.432	0.004	0.557
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.083	[.064,.104]	0.074	0.908	0.001	0.236
	CROSSDDI	0.557	[.525,.588]	0.576	0.427	0.001	0.561

Table 3: Full results including accuracy. Confidence intervals correspond to P.Int. Accuracy is shown for completeness but should not be interpreted as a reliable measure under PU evaluation.

Subset	Pairs	Share (%)	Positives	Recall (%)
Direct DrugBank match	558	55.80	465	85.81
No direct match	442	44.20	332	18.98

Table 4: Direct DrugBank coverage and end-to-end recall over confirmed positives in the randomly sampled 1,000-pair DDInter evaluation set.

G RAG Baseline Prompt

For transparency, we report the full prompt template used by the RAG baseline. The baseline receives the same retrieved DrugBank and PubMed evidence used by the verification paths. DrugBank evidence is formatted as up to three structured records, each containing an evidence ID, DrugBank interaction ID, counterpart name, retrieval rank bucket, severity, evidence level, and description. PubMed evidence is formatted as up to three retrieved articles, each containing an evidence ID, PMID, title, journal, publication date, and abstract. If no evidence is retrieved from a source, the corresponding section explicitly states that no records or documents were retrieved.

Prompt RAG: Interaction Verdict Generation.

```
You are an expert clinical pharmacologist
    evaluating a possible
drug-drug interaction.
Your task is to determine whether there is an
    interaction
between {drug_a} and {drug_b} based only on the
    retrieved
evidence below. Do not use outside knowledge.

DrugBank Evidence:
{drugbank_context}

PubMed Evidence:
{pubmed_context}

Instructions:
1. Use only the evidence provided above.
2. Classify the pair as "Interaction" if the
    evidence
   explicitly supports an interaction, mechanism,
       or
   clinical risk involving both drugs.
3. Classify the pair as "Non-Interaction" if the
     evidence
   explicitly states there is no interaction
       between the drugs.
4. Classify the pair as "Unknown" if the
    evidence is
   insufficient, indirect, conflicting, or does
       not explicitly
   establish either of the above.
5. If the evidence is mixed or unresolved,
    choose "Unknown".
6. Return valid JSON only.

Output schema:
{"rationale": "One or two sentences grounded in
    the retrieved
evidence", "final_label": "Interaction | Non-
    Interaction |
Unknown"}
```

GRAFT: Gated Retrieval-Augmented Fine-Tuning for Relation Extraction

Yuhang Jiang and Ramakanth Kavuluru
Division of Biomedical Informatics, Department of Internal Medicine
Department of Computer Science
University of Kentucky, Lexington, KY USA
yj38@iu.edu,ramakanth.kavuluru@uky.edu

Abstract

Even in the era of large language models (LLMs), biomedical relation extraction (RE) still plays a major role in timely creation of knowledge graphs that further guide biomedical knowledge discovery. The main task in RE is to extract a relation "as expressed" in an input text. At times, crucial definitional information or other auxiliary information about the entities involved may be missing from the input text. Augmenting it from other external textual sources appears helpful on the surface but can be harmful too, as these sources can overwhelm the signal in the original input, leading to false positives or false negatives. To counter this, we leverage a pre-trained biomedical text retriever to augment original inputs with additional instance-specific snippets. This is done through a gating mechanism that allows the retrieved snippets to enhance but not overwhelm the signal from the original input. We evaluate our approach on three standard biomedical relation extraction datasets (CDR, BioRED, and ChemProt) and show consistent improvements (up to 10 F1 points) compared with strong supervised baselines involving both encoder and decoder models. All our code and the datasets used are available for reuse: https://github.com/bionlproc/GRAFT-RE.

1 Introduction

Biomedical relation extraction (RE) is a core task in biomedical natural language processing (Huang et al., 2024), considering relations among biomedical entities drive etiology, pathology, treatment, and recovery in human diseases. Relations span the translational science spectrum from biological links (e.g., protein-protein interactions) to translational connections (e.g., gene-disease associations) and clinical information (e.g., drug-disease treatment relations). The goal in RE is to infer and classify the relation between a pair of input entities, typically from a fixed set of relation types. Relations extracted lead to knowledge graphs (KGs), which further enhance biomedical knowledge discovery via reasoning over KGs (Raghavan et al., 2021). Although LLMs have become general purpose NLP tools, KGs are still deemed essential to reduce hallucinations and ground reasoning in discovery tasks (Xiong et al., 2025; Xu et al., 2025). Hence, overall, RE remains an important primitive in BioNLP.

Consider the sentence: "We conclude that tamoxifen therapy is more effective for early stage breast cancer patients." From this sentence an RE model is expected to extract the relation (tamoxifen, TREATS, breast cancer) where tamoxifen is the *subject* entity, breast cancer is the *object* entity, and TREATS is the relation type or more formally the *predicate*. RE by definition implies that the only input provided is the sentence (or document) from which the relation is to be extracted. This means, we are **not** looking at general (global) knowledge about the entities mentioned in the input, but more interested in relations as explicitly stated (or implied) in the input (local assertions). But a given input often may not include background information about the entities needed to tease out a potential relation. This could be in the form of definitions (or glosses) of the entities in the input. This could also be in the form of hyponyms of entities involved. For the TREATS example we chose earlier, the definition of tamoxifen in the Medical Subject Heading (MeSH) vocabulary includes: "Tamoxifen acts as an anti-estrogen (inhibiting agent) in the mammary tissue, but as an estrogen (stimulating agent) in cholesterol metabolism, bone density, and cell proliferation in the endometrium." In the MeSH hierarchy we can also see that tamoxifen is a "benzylidene compound." In addition to the original text input to the RE, it might be beneficial to augment it with this extra contextual information. The contextual information can also be arbitrary pieces of text from the broad scientific literature that may enhance RE. Our current effort investi-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 920–931
July 3-4, 2026 ©2026 Association for Computational Linguistics

gates how to do this in a way that is both *effective* and *efficient*. Neither of these objectives is easy to meet because the extra text added as additional context to the input can confuse the model and drown the signal from the original input text. So in the end, the performance may actually worsen if the context contradicts what is stated in the input. The augmented context also increases the length of the input thus increasing computational cost.

Another concern in using additional context in RE may arise if the context already contains a directly stated positive relation, which is discussed in a complex and nuanced manner in the original input. In this case, a natural question to ask is if this constitutes cheating, considering the main goal is to extract relations as expressed in the input; the context might already contain the right answer. However, over the past decade encoder LMs already ingest billions of tokens of biomedical free text during pre-training (e.g., BioBERT (Lee et al., 2020) and BiomedBERT (Gu et al., 2021)). Furthermore, autoregressive LLMs are also pre-trained on PubMed abstracts and full text articles in PMC (Yang et al., 2022; Sallinen et al., 2025; Bolton et al., 2024). Given most modern RE efforts are almost exclusively based on encoders or LLM backbones, besides the input text that is explicitly provided, the models have entire PubMed/PMC at their disposal (via LM model weights), in an implicit manner. We propose to simply make this explicit along the lines of retrieval-augmented generation (RAG), already popular when using LLMs for question answering (QA) (Lewis et al., 2020; Xiong et al., 2024). The testing is always done against gold answers derived from the original input and hence the context's influence is strictly measured against what is available in the input.

In this paper, we present **G**ated **R**etrieval-**A**ugmented **F**ine-**T**uning (GRAFT) for RE, a parametrized approach that incorporates external retrieved context as part of the input for RE. GRAFT can seamlessly integrate with both small encoder models and autoregressive LLMs. It relies solely on a small encoder based, zero-shot retriever for biomedical RE (Jin et al., 2023), requiring no supervised data for retrieval. The documents retrieved are processed in parallel, thus reducing computational overhead significantly. Additionally, we incorporate a dynamic weighting (gating) layer that amplifies useful information from the context, while filtering out less relevant content. We evaluate GRAFT on three widely used biomedical RE datasets and show promising improvements across all of them. While this paper is focused on biomedicine, the method is agnostic of the domain and can be employed as long as there is a way to retrieve general Web text using well known models for the general domain (Izacard et al., 2022; Wang et al., 2022; Ni et al., 2022). The code and datasets used for this paper are available: `https://github.com/bionlproc/GRAFT-RE`.

2 Related Work

Several efforts on retrieval-augmented biomedical tasks focus on biomedical QA (Xiong et al., 2024; Sohn et al., 2024; Jeong et al., 2024; Das et al., 2024). Specifically, Xiong et al. (2024) benchmark a spectrum of retrieval–generator pairings across BioASQ, MedQA and PubMedQA revealing that naive coupling of powerful generators with off-the-shelf BM25 retrieval leaves substantial head-room that can be exploited by hybrid dense-sparse retrievers and cross-encoder rerankers. Sohn et al. (2024) extend this line of work with RAG^2, where rationale-guided query reformulation and a lightweight snippet filter reduce noise propagation and boost answer F1 by up to 6 points on three medical QA benchmarks.

Regarding other related efforts for RAG in biomedicine, Lopez et al. (2025) present CLEAR (Clinical Entity Augmented Retrieval), a task-aware RAG pipeline that first detects clinically salient entities in a potentially long clinical note, then restricts retrieval to context windows surrounding those entities, thereby trimming the context fed to the generator. This use case is different from GRAFT's use of external context to augment an input text for RE. BiomedRAG (Li et al., 2025) retrieves chunk-based documents from an external database and concatenates them directly into the LLM input, using a custom chunk scorer trained with a perplexity-driven supervision signal. It addresses noise at the retriever level (via better chunk selection), whereas GRAFT addresses it at the model level (gated fusion). Furthermore, GRAFT supports smaller encoder models for RE while BiomedRAG is an LLM-only method.

There are two related efforts in general domain RE. First, RetrievalRE (Chen et al., 2022) uses a non-parametric kNN retriever to select examples from the training data to be used in an interpolated manner along with a conventional parametric model's predictions. Hence, there is no

use of external knowledge as in GRAFT. Next, RAG4RE (Efeoglu and Paschke, 2025b) and its fine-tuned extension (Efeoglu and Paschke, 2025a) concatenate the retrieved context directly into the prompt and rely on LLMs, without any mechanism to control the influence of retrieved documents. In contrast, GRAFT processes retrieved documents in parallel, employs a learned gating mechanism to dynamically modulate retrieval influence, and is effective with small encoder models too. To summarize, BiomedRAG (Li et al., 2025) and RAG4RE (Efeoglu and Paschke, 2025a) differ from GRAFT in two aspects: (1) both target only LLM backbones, whereas GRAFT applies equally to small encoder models and causal LLMs. (2) GRAFT processes retrieved documents in parallel rather than concatenating them, decoupling per-document context length from the number of retrieved snippets.

Finally, we note that augmenting RE with auxiliary textual or structured information is a recurring theme in our prior work: Jiang and Kavuluru (2025) use instance-adapted predicate descriptions to enrich relation representations in a dual-encoder setup, while Jain et al. (2024a) cast document-level RE as context-guided link prediction and Jain et al. (2024b) incorporate structured knowledge for cross-document RE. GRAFT extends this direction by drawing on dynamically retrieved unstructured text rather than fixed predicate definitions or curated knowledge bases.

3 The GRAFT Method

3.1 Retriever model

To eliminate the need for retriever training, we optimize our entire system using MedCPT, a zero-shot biomedical IR model (Jin et al., 2023) trained on PubMed search logs by researchers at the US National Library of Medicine. In biomedical RE, the primary source of information lies in the entities themselves; entity traits can offer critical insights that inform the eventual relationship. Therefore, the retrieved information about the entities serves as ancillary context, aiding the prediction model in determining the relationship. We prepare two kinds of queries for MedCPT retrieval:

1. **Entity-wise query** to search for documents that describe or characterize an entity. We use the query *"What is the <entity_type> <entity_string>?"*. This is launched for both entities involved.
2. **Entity-pairwise query** to retrieve documents that discuss both entities and may reveal interactions between them. This query is formatted as: *"What is the interaction type between the <entity1_type> <entity1> and the <entity2_type> <entity2>?"*.

For biomedical RE, PubMed is an ideal choice for fetching snippets for augmentation and MedCPT comes pre-loaded with it. Additionally, we incorporate Wikipedia due to its broad complementary coverage of various biomedical concepts. Thus, the resulting documents retrieved constitute two components, the entity-wise documents and pairwise documents, both featuring PubMed abstracts and Wikipedia articles.

3.2 Zero-shot RE

To assess zero-shot performance in conjunction with the retrieved documents, we merge these documents with the input text for RE as part of the prompt. We test this strategy with the `Llama-3.2-Instruct-3B` model (Grattafiori et al., 2024), which allowed us to assess how effective the retrieved MedCPT documents are without training data. To ensure consistent outputs, we prompt the model to produce responses in a predefined JSON format, as shown in Figure 6 (of the Appendix). We present the results in Figure 1 for two datasets: chemical-disease relations (CDR (Li et al., 2016)) and chemical-protein relations (ChemProt (Krallinger et al., 2017)). We notice that the zero-shot prediction results remain suboptimal. The retrieved documents clearly hurt the performance as they likely introduced excessive noise when used in this simple way to augment the original input. The absence of clearly defined filtering mechanisms to distinguish between genuinely useful background information and irrelevant context may have caused the model to overlook or misinterpret cues essential for RE. Thus, simply appending external context can worsen results.

3.3 Fine-tuning GRAFT

To address the issues with naive augmentation, we propose using a fine-tuned model rather than a purely zero-shot approach. Although frontier LLMs can alleviate certain issues in zero-shot settings (Brokman et al., 2025), fine-tuning a smaller model offers a more cost-effective option. But several challenges remain unresolved: (1) incorporating extensive external documents into the trans-

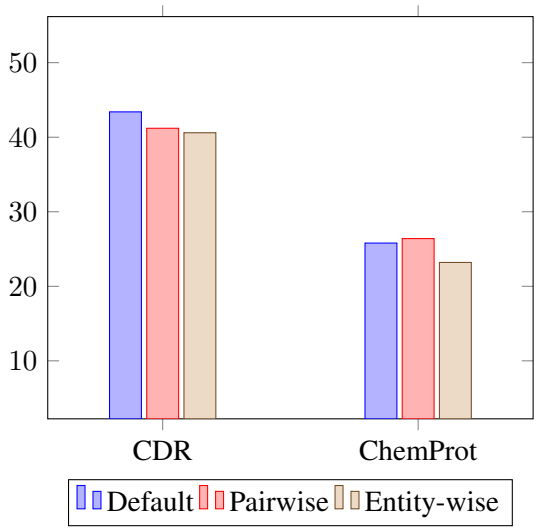


Figure 1: Zero-shot performance (in F1 scores) on CDR and ChemProt datasets using different types of retrieved documents. "Default" indicates that no external documents were applied. Up to 6 snippets were tried and the top score is reported.

former architecture's self-attention mechanism remains computationally expensive, incurring $O(n^2)$ computational complexity, and (2) failing to discard irrelevant documents still introduces additional noise into the model. The relation could be easily determined from the original context in some simple cases; the retrieved documents for these instances may lead to errors.

Therefore, instead of appending documents, we opt to input documents in parallel for two reasons: (1) we assume each document carries its own distinct information and should be processed independently, and (2) this reduces the computational overhead that would arise from cross-document interactions. As a result, the generative capability of language models is replaced by a BERT-like classification approach. Furthermore, inputting documents in parallel drastically reduces context length limitations, enabling the method to be applied to smaller models as well.

Specifically, let $\mathcal{M}$ be any transformer based model, x the input relation text, and $D = \{d_1, d_2, \ldots, d_n\}$ the set of retrieved documents. The model encodes the input text and each document independently:

$$h_x = \mathcal{M}_t(x), \quad h_i = \mathcal{M}_d(d_i), \ \forall i \in \{1, \ldots, n\}.$$

We form the representation by concatenating token-wise the text and document embeddings, then pass it through a lightweight, two-layer self-attention module to enable cross-interaction:

$$H_x = Self\,Attention([h_x \parallel h_1 \parallel h_2 \parallel \ldots \parallel h_n]).$$

Furthermore, we design a gating mechanism that allows less attention to unimportant documents or when the input text is very informative, needing less external context interference. It employs an MLP layer as the scorer to evaluate the original input text and assign a dynamic weight for the retrieved documents. We define this weight as

$$g_x = \sigma\left(MLP(h_x)\right) \in (0, 1).$$

Before we discuss how this weight g_x is incorporated into the training and inference process, we first ought to cover how the relation is represented. For bi-directional models, we explicitly mark each entity span with unique boundary tokens so the model can localize and differentiate the arguments of the relation. For example, the entities `acetaminophen` and `liver injury` are marked as: "`The analgesic [E1] acetaminophen [/E1] may induce [E2] liver injury [/E2] in susceptible patients.`" Let h_x^{E1} and h_x^{E2} be the contextualized output of `[E1]` and `[E2]`, the relation representation $Rep(h_x)$ in bi-directional model is defined as

$$Rep(h_x) = LayerNorm([h_x^{E1} \parallel h_x^{E2}]).$$

In causal (autoregressive) models, where each token attends solely to preceding tokens, we enclose the entities in boundary markers and append an additional `[EOT]` token at the end of the sequence, for the same example: "`The analgesic [E1] acetaminophen [/E1] may induce [E2] liver injury [/E2] in susceptible patients. [EOT].`" Let $h_x^{[EOT]}$ be the contextualized output of `[EOT]`, the $Rep(h_x)$ in causal model form is defined as

$$Rep(h_x) = LayerNorm(h_x^{[EOT]}).$$

With this setup for $Rep(h_x)$, we derive final relation labels via

$$\hat{h}_x = \mathbf{W_1}(Rep(h_x))$$

where $\mathbf{W_1}$ is the projection layer such that $\hat{h}_x$ is an r-dimensional vector with r being the total number of possible relations (label cardinality). Without the gating mechanism, the loss function is

$$\mathcal{L}_{\text{orig}} = -\sum_{j=1}^{r} y_j \log(\hat{h}_x^j).$$

As H_x is essentially h_x with retrieved documents concatenated, the notion of $Rep(\cdot)$ is applicable to H_x too. Therefore, the aggregated relation representation controlled by the weight g_x is

$$\hat{H}_x = (1-g_x)\mathbf{W_1}(Rep(h_x)) + g_x\mathbf{W_2}(Rep(H_x))$$

where $\mathbf{W_2}$ is a projection matrix just as $\mathbf{W_1}$. Now the aggregate loss function

$$\mathcal{L}_{\text{agg}} = -\sum_{j=1}^{r} y_j \log(\hat{H}_x^j),$$

and the full loss incorporates both the original and the aggregate losses (involving retrieved docs) into

$$\mathcal{L} = \mathcal{L}_{\text{agg}} + \mathcal{L}_{\text{orig}}.$$

During inference, the output logits ought to be informed by the augmentation of the retrieved documents. Hence, the aggregated form $\hat{H}_x$ is used to predict the relation.

4 Experiments

4.1 Datasets

We conduct the experiments on three commonly used biomedical relation extraction datasets: CDR, ChemProt and BioRED. Table 1 displays the resulting processed statistics of the datasets.

- **CDR** (Li et al., 2016) is annotated at the entity level for chemical-disease inducement relations of PubMed abstracts. Each entity is normalized to a unique identifier (e.g., *lidocaine* → *D008012*). Therefore, one identifier could sometimes refer to multiple entity strings. Because there are only two entity types: chemical and disease, and one relation type (inducement), the task simplifies to binary classification. To create a more difficult and realistic evaluation setting, we sever the linkage between mentions that share the same identifier, splitting them into separate relation instances. This adjustment ensures that each entity mention is handled independently, making the task more challenging than the standard setup that operates on normalized entities.

- **BioRED** (Luo et al., 2022) dataset is built on top of the original CDR abstracts, but it greatly expands both the entity and relation inventories, incorporating six entity types (chemical, disease, gene/protein, ...) and eight relation categories. Like CDR, BioRED includes normalized entity annotations, so we process it in the same manner as CDR dataset.

- **ChemProt** (Krallinger et al., 2017) captures interactions between chemicals and genes or proteins, collected from PubMed abstracts and spanning five relation types. Because the annotations are at the entity mention level, we apply no additional preprocessing.

Dataset	# Pred.	# Train	# Valid	# Test	Nor.
CDR	1	8441	8340	8502	Yes
BioRED	8	37286	11407	11880	Yes
ChemProt	5	17991	11328	15474	No

Table 1: Statistics of datasets used (columns 3–5 are numbers of instances/relations used in our experiments).

4.2 Retrieval corpora

Corpus	#Snippets	Ave. L	Domain	Entity-wise	Pairwise
PubMed	23.9M	296	Biomed.	✓	✓
Wikipedia	29.9M	162	General	✓	

Table 2: Statistics of the corpora; **Entity-wise** and **Pairwise** denote whether the corpus is applied with entity-wise query or pairwise query.

For retrieval, we employ two corpora, PubMed abstracts and Wikipedia articles, which together provide broad and diverse coverage of biomedical knowledge. PubMed serves as our biomedical corpus, whereas Wikipedia represents the general-domain corpus. As described in Section 3.1, we generate two retrieval queries for every entity pair: an entity-wise query, executed on both corpora, and an entity-pairwise query, applied only to PubMed abstracts due to efficiency considerations. The details of each corpus are in Table 2.

Dataset	#Pairs	#Mentions	%PubMed	%Wiki
CDR	21853	5384	69.4%	30.6%
BioRED	58998	6048	66.4%	33.6%
ChemProt	28449	12310	60.3%	39.7%

Table 3: Statistics on the retrieved snippets for all three datasets, combining the training, validation, and test splits; **#Pairs** and **#Mentions** denote the counts of unique pairs and unique mentions, respectively.

We query with each entity's surface form, so when multiple mentions share the same identifier

Model	*CDR*			*BioRED*			*ChemProt*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BiomedBERT	57.6	67.9	62.3	52.9	53.2	53.1	81.1	75.4	78.2
GRAFT w/ BiomedBERT (ours)	58.6	77.4	**66.7** ↑4.4	55.8	53.7	**54.7** ↑1.6	80.5	78.8	**79.6** ↑1.4
BERT-based									
Llama-3.2-3B-Instruct	69.6	62.0	65.6	54.2	51.6	52.9	73.4	59.9	66.0
w/ Retrieval	67.6	58.7	62.8	54.2	50.2	52.1	73.7	55.6	63.4
Llama-3.1-8B-Instruct	62.7	76.8	69.0	58.8	54.2	56.4	77.2	66.4	**71.4**
w/ Retrieval	67.4	71.2	69.2	59.2	55.1	57.1	77.2	65.4	70.8
GRAFT w/ Llama-3.2-3B-Instruct (ours)	67.9	74.7	**71.2** ↑5.6	64.7	62.7	**63.7** ↑10.8	76.6	65.6	70.6 ↑4.6
Causal LMs									

Table 4: Experimental results of GRAFT compared to corresponding baseline models. Our method consistently outperforms the corresponding baselines and surpasses the larger 8B model on CDR and BioRED. Improvements are shown with ↑, representing gains over the baseline using the same transformer model.

(e.g. *Parkinson's disease* and *PD*), every mention is submitted to the retriever separately. Following Xiong et al. (2024), we chunk each corpus into short snippets to enhance retrieval granularity, reduce noise, and ensure that the retriever returns passages focused on the target entities. These snippets are ranked by MedCPT according to the relevance scores it assigns; for efficient fine tuning, we select the top one to five snippets and feed them into GRAFT. Statistics of retrieved snippets are displayed in Table 3. Although Wikipedia holds more snippets overall, entity-wise queries return substantially more snippets from PubMed, reflecting our emphasis on biomedical tasks. Since the number of unique mentions is much smaller than that of unique pairs across all three datasets, using entity-wise queries for retrieval is considerably more efficient.

4.3 Baseline models

As outlined in Section 3.3, we evaluate our framework in two architectural settings: (i) bidirectional encoder models and (ii) left-to-right (causal) language models. For the causal-LM track, we adopt the recently released Llama-instruct models: `Llama-3.2-Instruct-3B` and `Llama-3.1-Instruct-8B`. For the encoder-based setting, we select BiomedBERT (Gu et al., 2021), a transformer encoder pre-trained exclusively on PubMed abstracts. We fine-tune BiomedBERT following Zhong and Chen (2021), a commonly used fine-tuning approach for RE tasks.

In baseline experiments, we fine-tune BiomedBERT using only the original text; its input-length limit of 512 tokens prevents appending any retrieved documents. (Note that this restriction applies only to naive augmentation; GRAFT allows for retrieval augmentation because of parallel processing of individual snippets.) For Llama models, we conduct two separate fine-tuning regimes: one on the raw input and one on document-augmented input, both optimized with the standard language modeling objective:

$$\mathcal{L}_{\mathrm{LM}}(\theta) = -\sum_{t=1}^{T} \log P_\theta\big(x_t \mid x_{<t}\big).$$

We employ the instruction template from Figure 6, consistent with the zero-shot configuration. We provide all model choices in Table 5.

Model	Size	Context Length	Domain
BiomedBERT	110M	512	Biomed.
Llama-3.2-Instruct	3B	128k	General
Llama-3.1-Instruct	8B	128k	General

Table 5: Model selection of our baseline experiments.

4.4 Main results

We present our results in two model categories: bidirectional encoders (e.g., BiomedBERT) and causal LMs. As shown in Table 4, GRAFT fine-tuned on BiomedBERT outperforms standard BiomedBERT fine-tuning, yielding F1 improvements of 1.4 % to 4.4 %. This demonstrates that even encoder models with limited context length

Model	*CDR*			*BioRED*			*ChemProt*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BiomedBERT *(baseline)*	57.6	67.9	62.3	52.9	53.2	53.1	81.1	75.4	78.2
GRAFT-Entity-wise	58.6	77.4	**66.7**	55.8	53.7	**54.7**	80.5	78.8	**79.6**
GRAFT-Pairwise	58.0	71.1	63.9	-	-	-	81.5	77.4	79.4
BERT-based									
Llama-3.2-3B-Instruct *(baseline)*	69.6	62.0	65.6	54.2	51.6	52.9	73.4	59.9	66.0
GRAFT-Entity-wise	67.9	74.7	**71.2**	64.7	62.7	**63.7**	76.6	65.6	**70.6**
GRAFT-Pairwise	62.7	76.8	69.0	-	-	-	77.4	64.8	**70.6**
Causal LMs									

Table 6: Results comparing different retrieval query criteria; BioRED results for pairwise snippets are excluded due to the vast number of unique entity pairs.

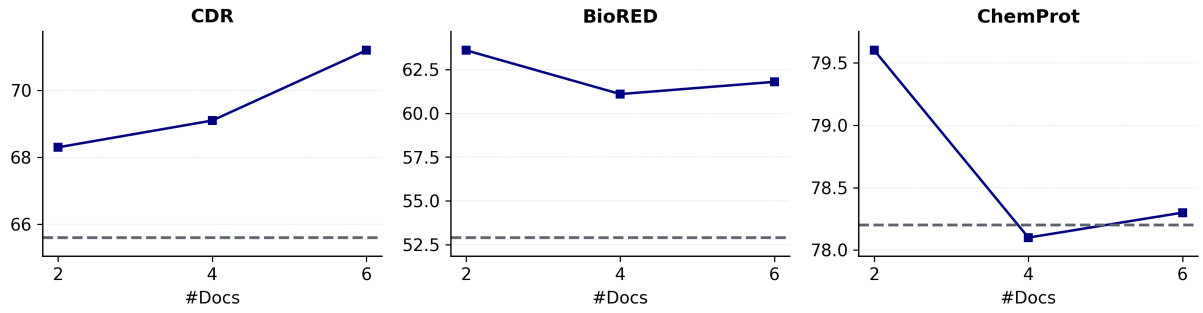


Figure 2: F1-score comparison based on different numbers of entity-wise documents retrieved.

can benefit from retrieval augmentation when fine-tuned with GRAFT. For causal LMs, we also evaluate a retrieval-augmented variant of vanilla fine-tuning to leverage their extended context capacity. GRAFT-tuned Llama-3.2-3B-Instruct outperforms the baseline on all three datasets, with F1 gains of 5.6%–10.8%. Additionally, GRAFT with the 3B-parameter Llama model surpasses a larger 8B-parameter Llama on the CDR and BioRED datasets by 2.0%–6.6% F1.

The magnitude of improvement varies across datasets and tracks the difficulty of the underlying task. ChemProt has a high baseline F1 in the BiomedBERT setting (78.2 vs. 62.3 for CDR and 53.1 for BioRED), leaving relatively less headroom for improvement; the corresponding 1.4 F1 gain with GRAFT likely reflects this ceiling effect. Conversely, BioRED, the most challenging dataset with eight relation types and the lowest baseline F1 (52.9 for Llama-3B, 53.1 for BiomedBERT), shows the largest gain (10.8 F1 with Llama-3B).

We observed an interesting phenomenon: standard fine-tuning of the 3B-parameter Llama model with retrieved documents actually degrades performance compared to fine-tuning without document augmentation. However, with the larger 8B-parameter model, this performance drop is alleviated, suggesting that the additional complexity from retrieved documents is more challenging for smaller LMs to manage. GRAFT enables smaller language models to benefit from retrieval augmentation, which is an advantage that, with naive concatenation, is only realized by larger architectures.

We note that the *w/Retrieval* rows of Table 4 are akin to the core RAG4RE design (Efeoglu and Paschke, 2025a): retrieved snippets are concatenated into the prompt and the model is fine-tuned with the standard language modeling objective, without any gating or noise-control mechanism. The gap between these rows and GRAFT rows isolates the contribution of gated parallel encoding rather than retrieval itself.

We also present results for various document-retrieval schemes in Table 6. As BioRED contains over 50K unique relation pairs, we did not evaluate pairwise retrieval on it. Across the other two

datasets, entity-wise queries consistently achieve higher performance for both BERT-based models and causal LMs. Since each dataset has fewer unique mentions than unique pairs, using entity-wise queries is inherently more efficient.

5 Discussion

5.1 Varying number of snippets

We vary the number of retrieved snippets/documents and evaluate how this affects performance. The results are shown in Figure 2. To ensure the efficient fine-tuning, we experiment with retrieving between 2 and 6 snippets, finding that all three datasets deliver strong results with just two retrieved snippets. Increasing the number of retrieved snippets led to performance declines on BioRED and ChemProt, implying that the inclusion of too much information can degrade performance. From a computational angle, limiting retrieval snippets also keeps both memory footprint and latency low. With fewer snippets retrieved, the model can more accurately assign attention, reducing the risk of overfitting to noisy context.

5.2 Ablation on corpus domains

We found that the choice of corpus domain has an impact on model performance. To quantify this, we performed ablation experiments, omitting one corpus at a time, and present the results in Figure 3. For BioRED and CDR, the mixed-corpus setup consistently achieved the highest average F1. In particular, augmenting PubMed with Wikipedia improved CDR by 3% F1 and BioRED by 2% F1, demonstrating that general-domain text can effectively complement biomedical literature. Conversely, relying solely on one domain proved insufficient for these tasks, confirming that a hybrid corpus strategy yields the most robust results. One reason could be because Wikipedia has general introductory exposition on many biomedical topics and hence could offer a clearer signal that is easy to process semantically compared with complicated jargon-loaded long sentences typically seen in PubMed abstracts. With ChemProt, which targets chemical–gene/protein interactions, we observe a slightly different pattern: PubMed-only retrieval delivers near-optimal performance. Adding Wikipedia incurs a 1% F1 drop. The results indicate that for datasets like ChemProt, where interactions are narrowly defined, domain-specific literature is more essential.

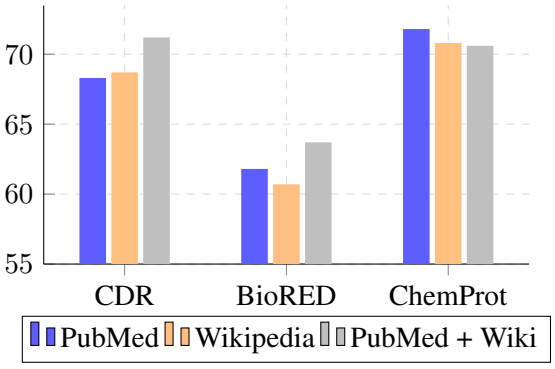


Figure 3: Model performance across different corpora.

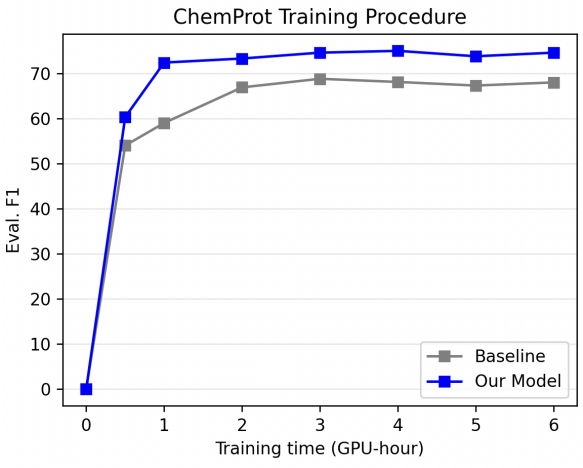
5.3 Computational efficiency

We report validation F1 scores on the ChemProt dataset across multiple epochs with our approach, comparing with the 3B-parameter Llama baseline model. As shown in Figure 4a, our model has reduced computational demands, and it converges more rapidly than the baseline, which relies on the standard language modeling objective. During inference, GRAFT requires only a single forward pass, making it faster than next-token generation methods, which require multiple forward passes. Meanwhile, when processing batched document inputs, GRAFT only requires a context window equal to the length of the longest individual document or text segment, whereas the baseline approach requires a context window large enough to accommodate all concatenated inputs. Consequently, GRAFT delivers substantially faster processing speeds than the baseline, which is shown in Figure 4b.

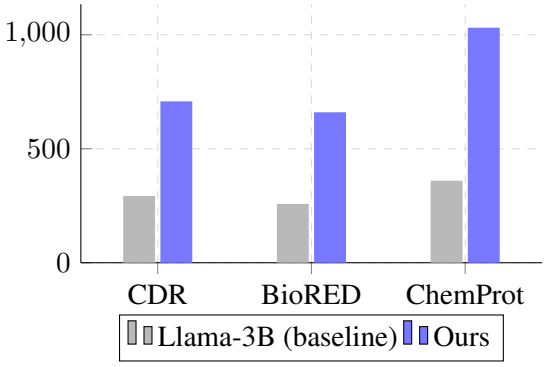
5.4 Case study: how much does prediction rely on retrieved information?

A natural concern with retrieval augmentation in RE is that retrieved snippets, drawn from corpora reflecting consensus knowledge about entity pairs, may bias the model toward predicting commonly attested relations rather than the relation as expressed in the input. This is especially relevant when the input asserts a novel, qualified, or context-specific relation that diverges from retrieved snippets. In Section 1, we posited that this is unlikely to happen since both training and testing are conducted against gold relations derived from the input.

Here to examine whether our approach addresses this concern, we look at the dynamic gate weights g_x across the entire CDR test set, shown in Figure 5 (y-axis: position in sorted order per g_x). Across



(a) Validation F1 over training epochs.



(b) Inference throughput (instances /min).

Figure 4: Computational performance: (a) learning curve and (b) processing speed.

all instances, g_x never exceeds 0.4. Thus, the gated combination consistently down-weights the retrieval branch. Despite this bounded influence, GRAFT improves CDR F1 from 62.3 to 66.7—a 4.4 F1 gain—indicating that retrieval supplies useful background cues without overwhelming local assertions in the input. This behavior is learned in GRAFT: the gate is trained jointly with the classifier and attenuates retrieval when the input is already informative.

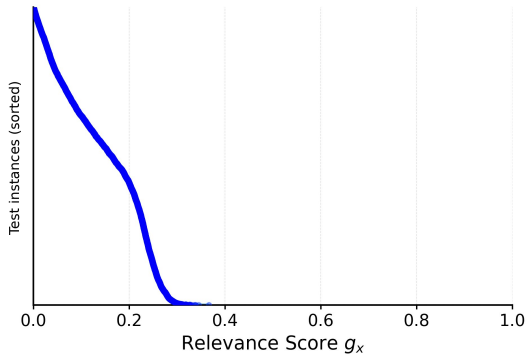


Figure 5: The gate g_x plotted (sorted) over instances of the CDR test set.

6 Conclusion

In this paper, we demonstrated that a training-free, zero-shot retrieval system can be effectively combined with task-aligned fine-tuning to address the challenges of lack of context in biomedical RE. Our proposed GRAFT framework seamlessly integrates a lightweight retriever (MedCPT) with both small encoder models (BiomedBERT) and causal language models (Llama), leveraging dynamic weighting to filter noise and amplify task-relevant information. Through extensive experiments on CDR, BioRED, and ChemProt, we showed that retrieval augmentation yields consistent F1 improvement, ranging from 1.4% to 4.4% for BiomedBERT and up to 10.8% for Llama-3B, while maintaining computational efficiency.

7 Limitations

Despite GRAFT's promise, our effort has some limitations and room for improvement. First, we evaluate on three biomedical RE datasets drawn from PubMed abstracts; generalization to clinical text or general-domain RE remains to be explored (although the method is agnostic to text type). Second, we use MedCPT as a frozen, zero-shot retriever; jointly optimizing the retriever for the RE objective could yield further improvements. Third, we tacitly assume adequate coverage of the target entities in the retrieval corpora; performance on rare or novel entities with sparse literature warrants further investigation. Finally, we use a fixed number of retrieved snippets per instance; adaptive retrieval strategies that dynamically adjust retrieval volume could offer additional gains.

Acknowledgment

This work is supported by the U.S. National Library of Medicine through grant R01LM013240. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. National Institutes of Health.

References

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Aviv Brokman, Xuguang Ai, Yuhang Jiang, Shashank Gupta, and Ramakanth Kavuluru. 2025. A benchmark for end-to-end zero-shot biomedical relation extraction with llms: experiments with openai models. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 44–55.

Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation extraction as open-book examination: Retrieval-enhanced prompt tuning. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2448.

Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, et al. 2024. Two-layer retrieval augmented generation framework for low-resource medical question-answering: proof of concept using reddit data. *arXiv preprint arXiv:2405.19519*.

Sefika Efeoglu and Adrian Paschke. 2025a. Fine-tuning large language models for relation extraction within a retrieval-augmented generation framework. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 1–7.

Sefika Efeoglu and Adrian Paschke. 2025b. Retrieval-augmented generation-based relation extraction. *Semantic Web*, 16(5):22104968251385519.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Ming-Siang Huang, Jen-Chieh Han, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2024. Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. *Briefings in Bioinformatics*, 25(3):bbae132.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. 2024a. Revisiting document-level relation extraction with context-guided link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18327–18335.

Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. 2024b. Knowledge-driven cross-document relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3787–3797.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Yuhang Jiang and Ramakanth Kavuluru. 2025. Relation extraction with instance-adapted predicate descriptions. In *Proceedings of the AMIA Annual Symposium*, pages 546–555.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, et al. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, Tahseen Rabbani, Trevor Brokowski, Meditron Medical Doctor Working Group, Tim G. J. Rudner, and Mary-Anne Hartley. 2025. Llama-3-meditron: An open-weight suite of medical LLMs based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2411.00300*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Guangzhi Xiong, Eric Xie, Corey Williams, et al. 2025. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7849–7857.

Ran Xu, Patrick Jiang, Linhao Luo, Cao Xiao, Adam Cross, Shirui Pan, Jimeng Sun, and Carl Yang. 2025. A survey on unifying large language models and knowledge graphs for biomedicine and healthcare. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6195–6205.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 50–61.

Appendix

Relation Extraction Template for LMs

Model Input: You are a helpful medical expert, and your task is to extract the relation between the given entity pairs using the relevant documents. The documents serve as a reference only; the primary focus for relation extraction should be on the input text. Organize your output in a json formatted as Dict{{"answer_choice": str(A/B/C/...)}}. Your responses will be used for research purposes only, so please have a definite answer.
Here are the relevant documents:
<Documents>
User: Here is the input text for relation extraction:
<Text>
What is the relationship between the subject <SUBJ> and the object <OBJ>?
Here are the potential choices: A. ... B. ... C. ... D. ... X. ...
Please generate your output in json.
Assistant:
Model Output: {"answer_choice": "X"}

Figure 6: An example prompt for RE with documents provided used for zero-shot methods.

Overview of the PsyDefDetect Shared Task at BioNLP 2026: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations

Hongbin Na[1,*] **Zimu Wang**[2,*] **Zhaoming Chen**[3] **Yining Hua**[4]
Rena Gao[5] **Kailai Yang**[6] **Ling Chen**[1] **Wei Wang**[2] **Shaoxiong Ji**[7,8]
John Torous[4] **Sophia Ananiadou**[6]

[1]University of Technology Sydney [2]Xi'an Jiaotong-Liverpool University
[3]University of Utah [4]Harvard University [5]The University of Melbourne
[6]The University of Manchester [7]ELLIS Institute Finland [8]University of Turku
Hongbin.Na@student.uts.edu.au, Zimu.Wang@liverpool.ac.uk

Abstract

We present an overview of PSYDEFDETECT, the shared task on detecting levels of psychological defense mechanisms in emotional support dialogues, co-located with BioNLP@ACL 2026. Grounded in the clinically validated Defense Mechanism Rating Scales (DMRS) framework, the task asks systems to classify a target seeker utterance, given its preceding dialogue context, into one of nine categories: seven hierarchical DMRS levels plus two auxiliary labels. Participants worked on PSYDEFCONV, a newly released corpus of 200 dialogues and 2,336 help-seeker utterances annotated under DMRS with substantial inter-annotator agreement. The task attracted 172 participants on CodaBench who produced 563 submissions, with 21 teams officially registering their results for the final ranking. The best system achieved a macro F1-score of 0.420, surpassing the strongest fine-tuned baseline reported in the dataset paper by a notable margin, yet leaving clear headroom. Our analysis highlights (i) a persistent tendency to over-predict the majority *High-Adaptive* class, (ii) a widening gap between accuracy and macro-F1 that reveals class-imbalance sensitivity, and (iii) the value of theory-aware and LLM-based approaches for fine-grained defensive-function classification. We release all task materials and invite the community to continue work on this novel intersection of clinical psychology and NLP.[1]

1 Introduction

Psychological defenses are, in Winnicott's words, the means by which the *False Self*, if successful in its function, hides the *True Self* (Winnicott, 2018). They are automatic strategies people use to regulate distress, and when used rigidly or excessively they are linked to poorer mental health outcomes and to interpersonal difficulties (Perry and Henry, 2004; Di Giuseppe et al., 2024). In emotional support conversations (ESC) (Liu et al., 2021), defenses shape what help-seekers disclose and how they accept or resist help—yet, despite rapid progress on empathy modeling (Hua et al., 2025b; Cai et al., 2024; Sorin et al., 2024), strategy selection (Kang et al., 2024; Hua et al., 2025a; Na et al., 2025), and affect understanding (Wang et al., 2024a, 2025; Ma et al., 2025; Zhao et al., 2025), the defensive function of seeker utterances remains largely unmodeled in current ESC systems (Di Giuseppe et al., 2024).

To catalyze research on this under-explored dimension of supportive dialogue, we organized **PSYDEFDETECT**, a shared task co-located with BioNLP@ACL 2026 on detecting *levels* of psychological defense mechanisms. The task is grounded in the clinically validated Defense Mechanism Rating Scales (DMRS) (Perry and Henry, 2004; Vaillant, 2012), and asks systems to classify each help-seeker utterance—given its preceding dialogue context—into one of nine categories: the seven hierarchical DMRS levels, augmented with two auxiliary labels for phatic and under-specified turns. The task is built on PSYDEFCONV (Na et al., 2026), the first conversational corpus annotated with DMRS-based defense levels. PSYDEFCONV is derived from ESConv (Liu et al., 2021) via stratified sampling over problem types and emotions, and was double-blind annotated to substantial agreement.

Participation. PSYDEFDETECT attracted strong community engagement. The task was hosted on CodaBench (Xu et al., 2022) from December 2025 through April 2026. Over the course of the evaluation period, the competition registered **172 participants** who together produced **563 submissions** to the leaderboard. By the final registration deadline, **21 teams** had officially submitted their results for inclusion in the ranking. The best system reached

[*]Co-leads of the shared task organization.

[1]Task website: https://psydefdetect-shared-task.github.io/; CodaBench: https://www.codabench.org/competitions/12124/.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 932–943
July 3-4, 2026 ©2026 Association for Computational Linguistics

a macro F1-score of 0.420, substantially surpassing the strongest fine-tuned baseline reported in the dataset paper (≈ 0.315 macro-F1 (Na et al., 2026)), while still leaving ample headroom for future research.

Contributions. This paper reports on the organization, dataset, participating systems, and results of PSYDEFDETECT. Our contributions are as follows:

- We introduce PSYDEFDETECT, the first shared task that operationalizes DMRS-based defensive functioning as an utterance-level classification problem over emotional support dialogue.
- We report task-level participation that reflects substantial community engagement, with 172 CodaBench participants, 563 leaderboard submissions, and 21 officially ranked teams.
- We present a methodological taxonomy of participating systems together with a full leaderboard analysis that benchmarks submissions against both zero-shot and fine-tuned baselines from the dataset paper.
- We characterize the dominant failure modes observed across the leaderboard, in particular the severe sensitivity to class imbalance and the systematic over-prediction of the High-Adaptive level, and use them to motivate concrete directions for future work at the intersection of clinical psychology and NLP.

2 Background and Related Work

Defense mechanisms and DMRS. Defense mechanisms are a cornerstone of psychodynamic theory (Freud, 1936). The *Defense Mechanism Rating Scales* (DMRS) organize roughly thirty mechanisms into seven hierarchical levels of defensive maturity, from *Action Defenses* (Level 1) to *High-Adaptive Defenses* (Level 7) (Perry and Henry, 2004; Vaillant, 2012). DMRS was originally designed for longitudinal clinical case formulation rather than for single-utterance classification; PSYDEFCONV (Na et al., 2026) adapts the scheme to conversational text by annotating at the *level* rather than the *mechanism* granularity, increasing identifiability and reliability.

Emotional support conversations. Research on ESC has built increasingly capable dialogue agents that alleviate user distress through multi-turn, strategy-grounded interaction, starting with the ESConv corpus (Liu et al., 2021) and extending to multi-strategy (Bai et al., 2025), synthetic (Zheng et al., 2023; Wang et al., 2024a), and reasoning-aware (Zhang et al., 2024) variants. These efforts have largely focused on *supporter* strategies, empathy, and affect modeling (Wang et al., 2024a; Cai et al., 2024; Sorin et al., 2024; Hua et al., 2025b; Xu et al., 2026), while the *defensive function* of seeker utterances has remained unmodeled.

Mental health classification. Prior shared tasks at BioNLP and CLPsych have addressed clinically-motivated classification problems such as suicide risk assessment, depression detection, and empathy prediction. BioNLP has covered a wide spectrum of medical-related challenges, such as clinical question-answering and summarization (Colelough et al., 2025; Soni et al., 2025; Xiao et al., 2025). However, tasks specifically addressing mental health have remained notably underrepresented. CLPsych shared tasks have addressed a range of mental-health-related problems, including depression and PTSD (Coppersmith et al., 2015), suicide risk assessment (Zirikly et al., 2019; Chim et al., 2024), online peer support (Milne et al., 2016), developmental mental-health prediction (Lynn et al., 2018), and longitudinal affective modeling (Tsakalidis et al., 2022; Tseriotou et al., 2025). However, these efforts have primarily centered on social media platforms such as X and Reddit, with comparatively limited attention to ESC. PSYDEFDETECT complements this line of work by introducing a theory-grounded, fine-grained classification problem that requires both local linguistic reasoning and context-aware interpretation.

3 Task Description

3.1 Problem Formulation

Given a multi-turn emotional support dialogue $D = (u_1, u_2, \ldots, u_t)$ between a help-*seeker* and a *supporter*, and a target help-seeker utterance u_t, the task is to predict a defense level label $y \in \mathcal{Y}$, where $\mathcal{Y}$ comprises the seven DMRS levels and two auxiliary labels:

- **Level 0 – No Defenses:** phatic or functional utterances that do not engage with psychological conflict.

- **Level 1 – Action Defenses:** passive aggression, help-rejecting complaining, acting out.
- **Level 2 – Major Image-Distorting:** splitting, projective identification.
- **Level 3 – Disavowal:** denial, rationalization, projection, autistic fantasy.
- **Level 4 – Minor Image-Distorting:** devaluation, idealization, omnipotence.
- **Level 5 – Neurotic:** repression, dissociation, reaction formation, displacement.
- **Level 6 – Obsessional:** isolation of affect, intellectualization, undoing.
- **Level 7 – High-Adaptive:** affiliation, altruism, anticipation, humor, self-assertion, self-observation, sublimation, suppression.
- **Level 8 – Needs More Information:** context is insufficient to assign a label with confidence.

Systems have access to the dialogue context preceding and including u_t, but must not use future turns—this mirrors the online nature of the clinical annotation setup (Na et al., 2026).

4 The PSYDEFCONV Dataset

4.1 Construction

PSYDEFCONV (Na et al., 2026) is built on top of ESConv (Liu et al., 2021). The organizers performed stratified sampling over the joint distribution of problem types and emotions in ESConv to obtain a representative 200-dialogue subset, annotated at the *seeker*-utterance level for defense functioning.

Two trained annotators with expertise in both psychology and NLP labeled each seeker turn independently and double-blind, reaching Cohen's $\kappa = 0.639$ (substantial agreement). Disagreements were adjudicated by consensus to form the gold standard. The annotation workflow was supported by DMRS CO-PILOT (Na et al., 2026), a four-stage LLM pipeline that produces stressor hypotheses, screens candidate DMRS items, validates evidence, and synthesizes ranked recommendations; it reduced mean annotation time by 24.0%.

Category	Total	Supporter	Seeker
# Dialogues	200	–	–
# Utterances	4,709	2,373	2,336
Avg. Turns per Dialogue	23.5 ± 6.6	11.9 ± 3.4	11.7 ± 3.3
Avg. Length of Utterances	19.8 ± 16.5	20.9 ± 17.0	18.8 ± 15.8

Table 1: Data statistics of PSYDEFCONV.

4.2 Corpus Statistics

Table 1 summarizes the basic statistics of PSYDEFCONV. The corpus contains 200 dialogues and 4,709 utterances, roughly evenly divided between supporter (2,373) and seeker (2,336) turns. Conversations are of moderate length, with a mean of 23.5 turns per dialogue and an average utterance length of 19.8 tokens. Only the 2,336 seeker turns carry DMRS defense-level annotations; supporter turns are provided as dialogue context but are not part of the evaluation.

4.3 Label Distribution

Table 2 reports the distribution of the 2,336 annotated seeker utterances across the nine defense labels, as well as their aggregation into four broader defensive categories. The distribution is strongly skewed toward the *High-Adaptive* level (Level 7), which alone accounts for 51.8% of all annotated utterances. Non-defensive or contextually ambiguous turns (Levels 0 and 8) jointly cover another 17.4%, leaving only roughly one third of the corpus for the remaining six defense levels. Within that tail, *Immature* defenses (Levels 1–4) aggregate to 18.9% while *Neurotic* defenses (Levels 5–6) account for 11.9%, with the individual minority levels ranging from 2.6% (Neurotic) to 9.2% (Obsessional).

This imbalance reflects the natural prevalence of defensive functioning in supportive dialogue (Na et al., 2026): adaptive coping such as self-assertion, self-observation, and seeking affiliation is, by design, what participants in ESC conversations most often express. However, the imbalance also means that systems optimized for overall accuracy can trivially favor the dominant class at the expense of minority defense levels. We adopt macro-averaged F1 over the positive classes (1–8) as the official ranking metric (Section 5) precisely to penalize such majority-class bias and to encourage systems that discriminate well across the full DMRS hierarchy.

Categories	Num	Proportion
Seeker's Defense Levels		
0 No Defenses	371	15.9%
1 Action Defenses	136	5.8%
2 Major Image-Distorting	77	3.3%
3 Disavowal	124	5.3%
4 Minor Image-Distorting	105	4.5%
5 Neurotic	61	2.6%
6 Obsessional	216	9.2%
7 High-Adaptive	1,211	51.8%
8 Needs More Information	35	1.5%
Overall	2,336	100.0%
Seeker's Defense Categories		
Non-Defensive/Ambiguous (0, 8)	406	17.4%
Mature Defenses (7)	1,211	51.8%
Neurotic Defenses (5, 6)	277	11.9%
Immature Defenses (1, 2, 3, 4)	442	18.9%
Overall	2,336	100.0%

Table 2: Distribution of the PSYDEFCONV dataset across annotated defense levels and aggregated defense categories.

5 Evaluation Setup

5.1 Platform and Protocol

The shared task was hosted on CodaBench (Xu et al., 2022), where participants downloaded the dataset, received starter baseline kits, and submitted predictions as zipped JSON files. Each team could register on the official leaderboard after completing the result-registration form.

5.2 Metrics

Following the dataset paper (Na et al., 2026), we report Accuracy, Macro Precision, Macro Recall, and Macro F1-score. Macro F1 evaluated over the positive classes (1–8) is the official ranking metric, chosen because it penalizes majority-class bias, weighs minority defensive levels equally with the dominant High-Adaptive level, and follows established practices for multi-class classification (Wang et al., 2022, 2024b; Chen et al., 2025). Level 0 is excluded from the official macro-F1 because it marks phatic or functional turns without defensive content. Level 8, by contrast, is retained because detecting that the available context is insufficient is an explicit task decision rather than a negative class. Furthermore, we include an additional leaderboard that aggregates performance across all classes, enabling a more comprehensive and holistic comparison of the submitted systems.

5.3 Baselines

The CodaBench page provided zero-shot prompting baselines using strong general-purpose LLMs. These baselines follow the same setup as the dataset paper; for reference, the strongest results reported in (Na et al., 2026) are as follows. Under zero-shot prompting, Gemini 2.5 Pro reaches an accuracy of 0.5636 and a macro-F1 of 0.2599, while DeepSeek-V3.2 with thinking enabled achieves comparable performance at 0.5572 accuracy and 0.2617 F1. Under supervised fine-tuning, the best results come from Ministral-8B (Acc = 0.6483, F1 = 0.3148) and InternLM3-8B (Acc = 0.6398, F1 = 0.3053).

6 Participating Systems

6.1 Participation Statistics

The PSYDEFDETECT shared task drew broad international engagement. The CodaBench competition registered **172 participants** and received **563 submissions** over the course of the evaluation period, with **21 teams** officially registering their results for the final ranking (Table 3). Of these, **15** accompanied their submission with a system description paper contributed to this volume. The author affiliations on these 15 papers span 12 countries—Australia, Bangladesh, Canada, China, Germany, India, Russia, Spain, Switzerland, the UK, the USA, and Vietnam—and 24 distinct institutions, including a number of cross-institution collaborations that combine academic NLP groups with clinical and biomedical informatics partners.

6.2 Methodological Taxonomy

The 15 described systems span a broad methodological space, but gravitate toward six recurring patterns, described below. Almost every team targets the same two task-specific pain points: the heavy skew toward Level 7 ("High-Adaptive"), and the ambiguity of DMRS boundaries for short, context-dependent utterances. Several teams combine two or more strategies, in which case we record the most distinctive contribution here.

(i) Multi-model ensembles and deliberative-agent architectures. Top-ranked systems reject the idea that a stronger single model can solve the task, and instead explicitly engineer *error independence* across voters or agents. NÜRNBERG NLP (ranked 1st; Steigerwald et al., 2026) constructs a 9-voter ensemble along three orthogonal axes—class granularity (9-class *gatekeeper* vs. 8-class *special-*

ists), training paradigm (generative vs. discriminative), and base model (Ministral vs. Phi-4)—with per-axis cross-validation folds. UTS (2nd; Galat and Rizoiu, 2026) operates a multi-phase deliberative council of Gemini 2.5 agents in which class-specific *advocates* rate evidence strength rather than vote, augmented by a targeted override ensemble of three fine-tuned Qwen-family models. TONI-NLP (9th; Paul et al., 2026) likewise finds that ensembles outperform any individual prompting, fine-tuning, or embedding-classifier baseline.

(ii) Retrieval- and rubric-grounded LLM classification. A second cluster of systems brings the DMRS clinical rubric into the prompt, either dynamically via retrieval or statically via curated boundary cues. PERCEPTIONLAB (3rd; Fahim et al., 2026) pairs dynamic DMRS-Q item retrieval (Gemini 2.5 Pro) with a Gemini 2.5 Flash classifier fine-tuned on reasoning traces distilled from the same Pro model, explicitly targeting "LLM polarization" toward extreme labels. DAL TEAM (10th; Chu et al., 2026) runs a retrieval-augmented LLM pipeline that decomposes prediction into a coarse-to-fine hierarchy and adds summary-based distillation of dialogue context. ZZUCS (6th; Huang et al., 2026) uses its *CoR-QLoRA* approach to encode task contracts, taxonomy definitions, and adjacent-level boundary cues into the prompts used for 8B QLoRA adaptation.

(iii) Parameter-efficient LLM fine-tuning. Most mid-to-top-ranked systems converge on QLoRA or equivalent PEFT on mid-scale open LLMs. LINGUIUTICS (4th; Adib et al., 2026) QLoRA-fine-tunes Qwen3-8B with grouped stratified cross-validation, minority-class lexical augmentation, and a post-hoc logit-bias calibration step that substantially improves the rare "Needs More Information" class. ERASERHEAD (7th; Horaira et al., 2026) fine-tunes Qwen3-14B with clinically informed prompts and iteratively retuned per-class oversampling. TRANSFORMER_1376 (12th; Saha et al., 2026) reformulates the task as conditional text generation and applies 4-bit QLoRA to Gemma-2-2B.

(iv) Encoder-only and domain-specific transformer fine-tuning. Three teams explore encoder backbones. NEURAL NEXUS (11th; Basu, 2026) fine-tunes ROBERTA-BASE with a composite objective combining focal loss, label smoothing, and square-root-dampened class weights, using role-tagged dialogue history and a `[TARGET]` marker on the target utterance. CS_METRO (15th; Rebayet et al., 2026) builds a three-stage pipeline of LLM-based dialogue summarization, domain-specific transformer fine-tuning (including MentalBERT and Mental-RoBERTa variants), and rule-based ensembling; they report that domain-specific encoders outperform generic LLM fine-tuning on this clinical task. KCL-COGSTACK (17th; Agarwal et al., 2026) contrasts flat fine-tuning, few-shot prompting, and a hierarchical coarse-to-fine classifier that exploits the DMRS label tree, finding the hierarchical design the most effective of the three.

(v) Synthetic data and theory-informed augmentation. VISHC (13th; Vu et al., 2026) focuses on data scarcity rather than model scale, proposing stressor-anchored, theory-driven synthetic data generation combined with a hybrid model that fuses language representations with structured clinical features. Lexical augmentation and oversampling are also used as secondary ingredients by LINGUIUTICS, ZZUCS, TRANSFORMER_1376, and ERASERHEAD.

(vi) Systematic exploration and negative-result studies. Two teams emphasize breadth of comparison over a single headline system. ALIEN-ANNOTATORS (19th; Karip and Hossain, 2026) systematically evaluates six open-source small language models ($\leq$ 9B) under zero-shot and fine-tuning regimes, finding that clinically grounded prompts consistently outperform bare label definitions and that model scale alone does not help zero-shot performance; their post-submission configuration (fine-tuning with 5-fold CV and logit averaging) reaches macro-F1 = 0.346, more than double their official submission. EXPLAINATORS (20th; Babakova et al., 2026) similarly runs four exploratory tracks—direct prompting, encoder fine-tuning, novel "state-of-mind" generation, and LLM fine-tuning—across DeepSeek-V3.2, Qwen-family models, and GLM-series models.

7 Results

7.1 Official Leaderboard

Table 3 lists the final leaderboard, ranked by macro F1-score on the held-out test set. The top system (Nürnberg NLP, F1 = 0.4200) outperforms the second-ranked team (UTS, F1 = 0.4055) by a small margin, while the overall spread of F1 scores is wide (0.063–0.420), underscoring the difficulty of

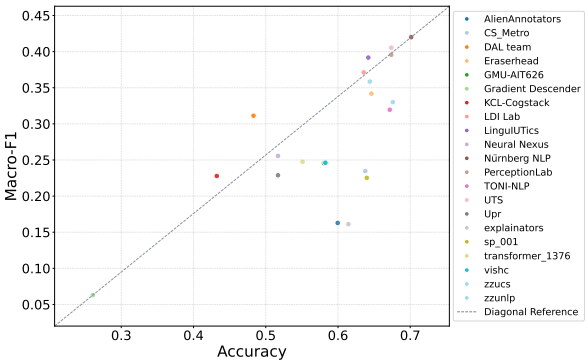


Figure 1: Scatter plot illustrating the relationship between accuracy and macro F1-scores across all submitted systems.

the task.

7.2 Key Observations

Top systems substantially outperform established baselines. The best-performing system, introduced by Nürnberg NLP, achieves a macro F1-score of 0.420, surpassing the best fine-tuned baseline reported in the dataset paper (Ministral-8B, F1 = 0.315) by approximately **10.5 absolute points**, and the best zero-shot baseline by **16 points**. These gains suggest that participants were able to leverage additional task-specific signals beyond standard fine-tuning setups. Nevertheless, the overall performance ceiling remains modest in absolute terms, indicating that the task continues to pose significant challenges.

Wide variability in system performance. System performance exhibits a wide dispersion, with macro F1-scores ranging from 0.063 (GRADIENT DESCENDER) to 0.420 (NÜRNBERG NLP), corresponding to a $6.7\times$ difference. The distribution further reveals a clear quartile structure: the top quartile achieves F1-scores of ≥ 0.37, the median lies ≈ 0.25, and the bottom quartile falls ≤ 0.23. This pronounced spread indicates that, despite the intrinsic difficulty of the task, there remains considerable room for methodological differentiation and performance gains across approaches.

Accuracy is a misleading proxy for F1-score. Several submissions attain relatively high accuracy while exhibiting substantially lower F1-scores, as shown in Figure 1. For instance, ZZUNLP (Acc = 0.676, F1 = 0.330) and TONI-NLP (Acc = 0.672, F1 = 0.320) demonstrate this discrepancy. This mismatch stems from the pronounced class imbalance, with the Level 7 majority class accounting for

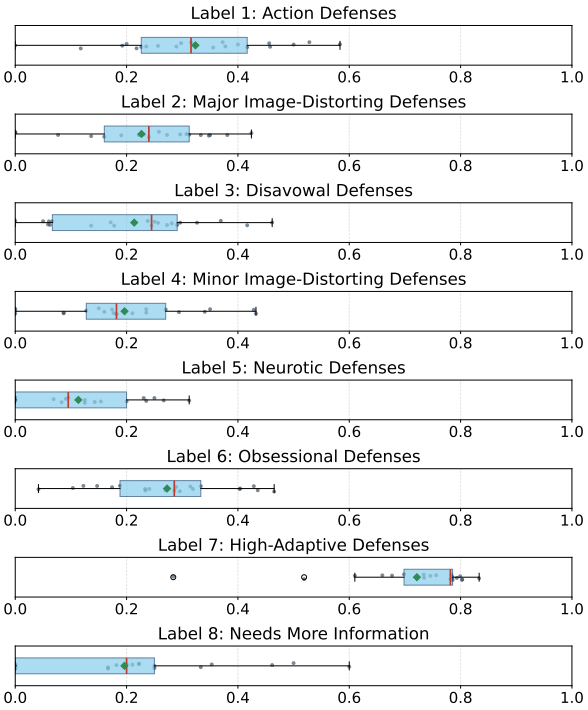


Figure 2: Per-class F1-scores across all submitted systems for the positive classes 1–8. Each box summarizes, for a given defense level, the distribution of F1 values over the 21 leaderboard submissions.

$\approx$52% of the dataset. Models that over-predict the High-Adaptive class can achieve competitive accuracy, but suffer from poor recall on minority classes, thereby degrading macro F1-scores. In contrast, systems such as NÜRNBERG NLP, PERCEPTIONLAB, and LINGUIUTICS achieve a more favorable balance, maintaining strong accuracy alongside improved recall across all classes.

8 Analysis and Discussion

8.1 Per-Class Performance

Figure 2 summarizes per-class F1 across the 21 submitted systems. Level 7 (High-Adaptive) stands apart: its median F1 is by far the highest and the inter-team spread is narrow, reflecting both its majority status ($\approx$52% of the corpus) and the distinctiveness of its surface markers. Every other class has a low median and a wide spread—widest for Levels 2–4, whose linguistic realizations overlap in affective-blame and reality-distortion cues, and for Level 8 (Needs More Information), on which a non-trivial fraction of systems collapse to zero. The narrow box for Level 7 marks a class where surface cues suffice; the wide boxes (Levels 1–6) mark classes where methodological choices translate directly into outcome differences.

			Positive classes (1–8)			All classes (0–8)		
Rank	**Team**	**Acc**	**P**	**R**	**F1**	**P**	**R**	**F1**
1	Nürnberg NLP	0.7013	0.4510	0.4036	0.4200	0.4959	0.4639	**0.4732**
2	UTS	0.6737	0.4607	0.3884	0.4055	0.4915	0.4327	0.4450
3	PerceptionLab	0.6737	0.4263	0.4089	0.3956	0.4721	0.4479	0.4402
4	LinguIUTics	0.6419	0.4003	0.3958	0.3917	0.4416	0.4570	0.4427
5	LDI Lab	0.6356	0.3770	0.3892	0.3713	0.4277	0.4422	0.4244
6	zzucs	0.6441	0.3969	0.3520	0.3585	0.4402	0.4166	0.4135
7	Eraserhead	0.6462	0.4075	0.3193	0.3418	0.4482	0.3801	0.3947
8	zzunlp	0.6758	0.4991	0.2891	0.3300	0.5381	0.3578	0.3909
9	TONI-NLP	0.6716	0.4701	0.2839	0.3196	0.5094	0.3560	0.3813
10	DAL team	0.4831	0.4187	0.2773	0.3113	0.4165	0.3516	0.3391
11	Neural Nexus	0.5169	0.2480	0.2867	0.2556	0.3140	0.3259	0.3080
12	transformer_1376	0.5508	0.2669	0.2351	0.2475	0.3132	0.2890	0.2979
13	VISHC	0.5826	0.2588	0.2503	0.2462	0.3168	0.3069	0.3045
14	GMU-AIT626	0.5805	0.2844	0.2328	0.2455	0.3298	0.2839	0.2952
15	CS_Metro	0.6377	0.3001	0.2572	0.2346	0.3673	0.3131	0.3004
16	Uprm	0.5169	0.2600	0.2161	0.2288	0.3081	0.2854	0.2877
17	KCL-Cogstack	0.4322	0.2913	0.2517	0.2278	0.3359	0.3171	0.2868
18	sp_001	0.6398	0.3057	0.2309	0.2253	0.3617	0.3060	0.2953
19	AlienAnnotators	0.5996	0.1555	0.1975	0.1628	0.2027	0.2823	0.2251
20	explainators	0.6144	0.2366	0.1660	0.1612	0.3041	0.2275	0.2296
21	Gradient Descender	0.2606	0.1383	0.0455	0.0629	0.1465	0.1501	0.0946
Baseline – Ministral-8B (fine-tuned) (Na et al., 2026)		0.6483	0.3397	0.3045	0.3148	0.3978	0.3640	0.3745
Baseline – Gemini 2.5 Pro (zero-shot) (Na et al., 2026)		0.5636	0.2749	0.2612	0.2599	0.2907	0.3107	0.2893

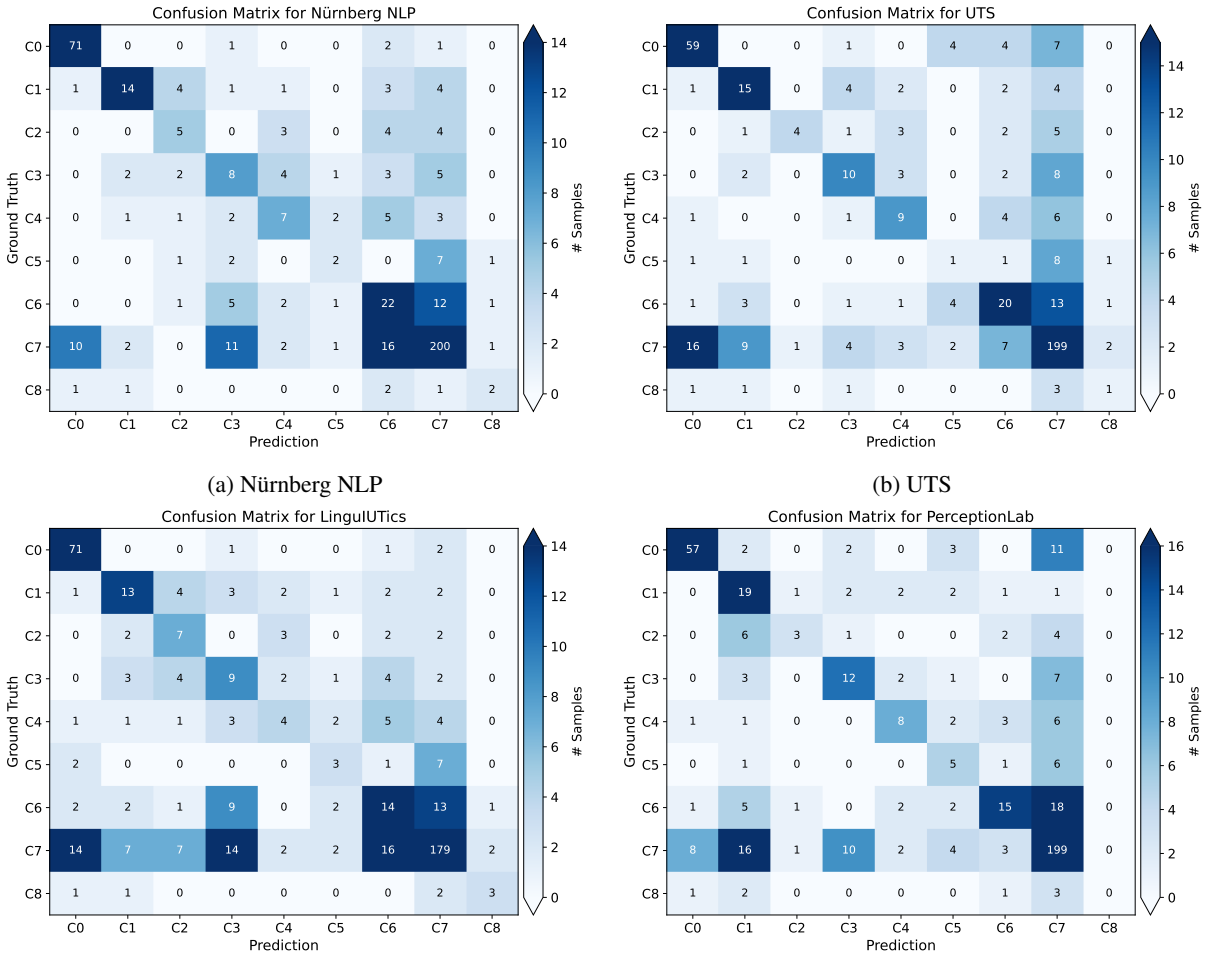
Table 3: Official PSYDEFDETECT leaderboard. Systems are ranked by macro-F1 over the *positive* classes 1–8 (the official metric of the shared task); for reference we additionally report macro-averaged precision, recall, and F1 computed over *all* nine classes (0–8). Accuracy is shared across both metric groups. The winning system's official macro-F1 of 0.4200 exceeds the strongest fine-tuned baseline reported in the dataset paper by $\approx$10.5 absolute points; its all-class F1 of 0.4732 exceeds that baseline by $\approx$9.9 points.

8.2 Error Patterns

Figure 3 shows confusion matrices for the four top-ranked systems. A single error pattern dominates all four: systematic *over-prediction of the High-Adaptive level* (C7), the "L7 attractor" identified by the dataset paper as the central failure mode of language models on PSYDEFCONV (Na et al., 2026). More interesting is *how* the four systems trade this attractor against minority-class recall. LINGUIUTICS is the most conservative on C7 (diagonal of 179 vs. $\approx$200 for the others), a direct effect of its minority-class augmentation and logit-bias calibration; the cost is some C7 recall, but the benefit is improved recovery of Level 8 instances. PERCEPTIONLAB shows a different signature: its dynamic DMRS-Q retrieval produces markedly higher Level 1 recall (19/28), while also exhibiting a wider over-prediction of C7. No single design choice resolves the attractor; different strategies trade one form of error against another, and the choice determines *which* minority classes a system can recover.

8.3 What Worked, What Did Not

Three takeaways emerge across the 15 system papers. First, *ensembling is the strongest differentiator at the top*: the two highest-scoring systems both engineer error independence explicitly—NÜRNBERG NLP along three orthogonal axes, and UTS through a deliberative agent council—while TONI-NLP (Paul et al., 2026) reaches the same conclusion through portfolio comparison. Second, *task-specific supervision beats backbone scale*: mid-scale backbones (Qwen3-8B, Ministral-8B, Qwen3-14B) dominate the top half of the leaderboard, and ALIENANNOTATORS (Karip and Hossain, 2026) report directly that scale alone does not reliably help zero-shot performance whereas clinically grounded prompts consistently do. Third, *zero-shot prompting alone is insufficient*: no system that relied primarily on zero-shot prompting of a large LLM exceeded F1 = 0.30, and every described top-10 system combines prompting with fine-tuning, retrieval, or ensembling. Clinical theory, packed into a prompt, is necessary but must be coupled with task-specific supervision to translate into accurate classification.



(a) Nürnberg NLP

(b) UTS

(c) LinguIUTics

(d) PerceptionLab

Figure 3: Confusion matrices for the four top-ranked systems on the held-out test set. Rows denote ground-truth labels, columns denote predictions. Color scales are clipped for readability; cell annotations show raw counts.

9 Future Directions

Four directions stand out as natural next steps. (1) *Dialogue-level defensive trajectories*: moving from utterance-level classification to modeling how defenses evolve across a conversation, building on the emotion-level trajectory analysis already included in the dataset paper (Na et al., 2026). (2) *Multilingual* PSYDEFCONV: annotating comparable corpora in Chinese, Japanese, Spanish, and other languages, with DMRS CO-PILOT lowering the clinical-annotation cost that has historically blocked cross-lingual scaling. (3) *Coupling defenses and supporter strategies*: jointly modeling how supporters adapt their strategies to seeker defensive functioning, closing the loop back to the broader ESC research programme. (4) *Theory-aware generation*: using defense-level predictions to condition emotionally supportive responses, operationalizing the seeker-model → supporter-response direction outlined as future work in the dataset paper. The first and fourth directions, in particular, turn PSYDEFDETECT from a static classification benchmark into a building block for clinically informed dialogue systems.

10 Conclusion

We introduced PSYDEFDETECT, the first shared task on detecting DMRS-grounded psychological defense levels in emotional support dialogues. The task drew 172 CodaBench participants, 21 ranked teams, and 15 system description papers across 12 countries. The winning system reached macro-F1 = 0.420, a ≈10.5-point gain over the strongest fine-tuned baseline, yet leaves clear headroom on minority classes where the "L7 attractor" dominates. We release all task materials to support continued work at the intersection of clinical psychology and NLP.

Limitations

This shared task has several limitations. First, PSYDEFCONV is a relatively small corpus derived from English ESConv dialogues, so results may not generalize to other languages, cultures, clinical settings, or naturally occurring therapy conversations. Second, the task labels defenses at the DMRS level rather than at the individual-mechanism level. This choice improves annotation reliability, but it also collapses clinically meaningful distinctions within each level. Third, defense interpretation remains inherently contextual: even with trained annotators and adjudication, short seeker utterances can be ambiguous without broader personal or longitudinal information. Finally, our methodological analysis is based on the 15 teams that submitted system description papers; leaderboard entries without accompanying papers are included in the quantitative ranking but cannot be analyzed in the same level of detail.

Acknowledgments

We thank all 21 participating teams and the 15 teams that contributed system description papers for making this shared task a success. We thank the BioNLP workshop organizers for hosting PSYDEFDETECT. We gratefully acknowledge the *Label Studio Academic Program* for providing access to the annotation platform used to construct PSYDEFCONV.

Ethical Considerations

Data. PSYDEFCONV is derived from ESConv (Liu et al., 2021) under its stated terms. Only derived annotations, dialogue identifiers, and code are redistributed; downstream users must obtain ESConv separately and comply with its license.

Intended use. The dataset and resulting models are intended *solely for research* on language and defensive functioning. They are **not diagnostic tools** and must not be used to make clinical, legal, or employment decisions about individuals. The corpus is in English and reflects the domains covered by ESConv, so results may not generalize across cultures or clinical settings.

References

Shefayat E Shams Adib, Ahmed Alfey Sani, Md Hasibur Rahman Alif, and Ajwad Abrar. 2026. LinguIUTics at PsyDefDetect: Iterative imbalance-aware finetuning of Qwen3-8B for psychological defense mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Shubham Agarwal, Thomas Searle, and Richard Dobson. 2026. KCL-Cogstack at PsyDefDetect: A hierarchical approach to detecting defense mechanisms in supportive dialogue. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Liudmila Babakova, Christopher Luongo-Vázquez, and Ilia Stepin. 2026. Explainators at PsyDefDetect: Hierarchical prompting and representation-based classification for psychological defenses. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Xin Bai, Guanyi Chen, Tingting He, Chenlian Zhou, and Yu Liu. 2025. Emotional supporters often use multiple strategies in a single turn. *Preprint*, arXiv:2505.15316.

Subhrajyoti Basu. 2026. Neural nexus at PsyDefDetect: Fine-tuning RoBERTa with focal loss and role-tagged dialogue history for defense level detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Torino, Italia. ELRA and ICCL.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. MedFact: A large-scale Chinese dataset for evidence-based medical fact-checking of LLM responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32340–32353, Suzhou, China. Association for Computational Linguistics.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

Anh Phuong Chu, Luong Duc Tran, Dat Hoang Do, Phuong Tu Mai, Quynh Hoang Le, and Cat Duy

Can. 2026. DAL team at PsyDefDetect: From supervised encoders to hierarchical LLM-RAG for psychological defense detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Brandon Colelough, Davis Bartels, and Dina Demner-Fushman. 2025. Overview of the ClinIQLink 2025 shared task on medical question-answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 378–387, Vienna, Austria. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Mariagrazia Di Giuseppe, Katie Aafjes-van Doorn, Vera Békés, Bernard S Gorman, Karl Stukenberg, and Sherwood Waldron. 2024. Therapists' defense use impacts their patients' defensive functioning: a systematic case study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 27(2):797.

Tamjid Hasan Fahim, Syed Asif Johan, and Saad Bin Maksud. 2026. PerceptionLab at PsyDefDetect: Overcoming LLM polarization via rubric-grounded retrieval and supervised clinical reasoning distillation for fine-grained ordinal classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Sigmund Freud. 1936. Inhibitions, symptoms and anxiety. *The Psychoanalytic Quarterly*, 5(1):1–28.

Dima Galat and Marian-Andrei Rizoiu. 2026. UTS at PsyDefDetect: Multi-agent councils and absence-based reasoning for defense mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Muhammad Abu Horaira, Mehreen Rahman, and Nahian Chowdhury. 2026. Eraserhead at PsyDefDetect: Prompt design and class rebalancing for psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025a. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230.

Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025b. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3):383–394.

Bin Huang, Liuyuan Su, Kaixuan Yuan, Guanghui Zhao, Shixin Zhang, and Kunli Zhang. 2026. zzucs at PsyDefDetect: Bridging long-tail imbalance and clinical rubrics for DMRS defense-level detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Siam Rahman Karip and Nahid Hossain. 2026. Alien-Annotators at PsyDefDetect: What lies between the lines: Probing lightweight open-source LLMs for psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Durjoy C. Paul, Arshitha Basavaraj, Callum Chan, Veronica Perez-Rosas, Diana Inkpen, Francisco Pereira, and Juan Antonio Lossio-Ventura. 2026. TONI-NLP at PsyDefDetect: Defense mechanism detection via LLM-based ensemble methods. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

J Christopher Perry and Melissa Henry. 2004. Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. *Advances in psychology*, 136:165–192.

Oarisa Rebayet, Radiul Walee, Symom Hossain Shohan, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2026. CS_Metro at PsyDefDetect: Detecting psychological defense mechanisms in mental health dialogues with summarization-enhanced transformer ensembles. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Pritha Saha, Shuvodwip Saha, and Anik Mahmud Shanto. 2026. transformer_1376 at PsyDefDetect: A QLoRA-based generative framework for context-aware psychological defense mechanism detection. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Sarvesh Soni, Soumya Gayen, and Dina Demner-Fushman. 2025. Overview of the ArchEHR-QA 2025 shared task on grounded question answering from electronic health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 396–405, Vienna, Austria. Association for Computational Linguistics.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *J Med Internet Res*, 26:e52597.

Philipp Steigerwald, Eric Rudolph, and Jens Albrecht. 2026. Nürnberg NLP at PsyDefDetect: Multi-axis voter ensembles for psychological defence mechanism classification. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

George E Vaillant. 2012. *Adaptation to life*. Harvard University Press.

Hoang-Thuy-Duong Vu, Quoc-Cuong Pham, and Huy-Hieu Pham. 2026. VISHC at PsyDefDetect: Mitigating data scarcity in psychological defense classification with context-aware synthetic augmentation. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, CA, USA. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024b. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Donald W Winnicott. 2018. Ego distortion in terms of true and false self. In *The person who is me*, pages 7–22. Routledge.

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. 2025. Overview of the BioLaySumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 365–377, Vienna, Austria. Association for Computational Linguistics.

Qingyang Xu, Yaling Shen, Stephanie Fong, Zimu Wang, Yiwen Jiang, Xiangyu Zhao, Jiahe Liu, Zhongxing Xu, Vincent Lee, and Zongyuan Ge. 2026. Do no harm: Exposing hidden vulnerabilities of LLMs via persona-based client simulation attack in psychological counseling. *Preprint*, arXiv:2604.04842.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal LLM for depression detection by integrating visual understanding into audio language models. *Preprint*, arXiv:2511.19877.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

SCoPE: Planning for Hybrid Querying over Clinical Trial Data

[1]Suparno Roy Chowdhury* [1]Manan Roy Choudhury* [1]Tejas Anvekar
[2]Muhammad Ali Khan [2]Kaneez Zahra Rubab Khakwani [2]Mohamad Bassam Sonbol
[2]Irbaz Bin Riaz† [1]Vivek Gupta†
[1]Arizona State University [2]Mayo Clinic
{srchowd3,mroycho1,tanvekar,vgupt140}@asu.edu
{khan.muhammad2,Khakwani.kaneezzahra,Sonbol.mohamad,riaz.irbaz}@mayo.edu

Project Page **Project Code**

Abstract

Systematic reviews of clinical trials require analysts to extract attributes that are rarely stored as ready-made columns. For example, the drug class of an immunotherapy named in a regimen, the additional agents combined with it, or whether a listed endpoint is a primary or secondary outcome. These attributes must be inferred from the visible content of other fields through normalization, classification, or structured extraction, and existing approaches such as direct LLM prompting, text-to-SQL, and agentic pipelines leave this reasoning implicit in a single generation step or pay a heavy execution cost for limited accuracy gains. We propose SCoPE (***S**tructured **C**linical hybrid **P**lanning for **E**vidence retrieval in clinical trials*), a multi-LLM planner-based framework that decomposes the task into row selection, structured planning, and execution. The planner makes the source field, reasoning rules, and output constraints explicit before answer generation, reducing ambiguity relative to direct prompting. We evaluate SCoPE on 1,500 hybrid reasoning questions over oncology clinical-trial tables against zero-shot, few-shot, chain-of-thought, TableGPT2, BlendSQL, and EHRAgent. Results show that explicit multi-LLM planning improves accuracy for reasoning-based questions while offering a stronger accuracy-efficiency tradeoff than heavier agentic baselines. Our findings position clinical trial reasoning as a distinct table understanding problem and highlight hybrid planner-based decomposition as an effective solution.

Question: For Pembrolizumab trials list **additional agents** in the treatment regimen beyond the ICI.

NCT	Treatment Regiment	Added Agents
NCT0203	Pembrolizumab+**Pemetrexed**	["pemetrexed"]
NCT0204	Pembrolizumab+**Pemetrexed +Carboplatin**	["pemetrexed", "carboplatin"]
NCT0206	Pembrolizumab+**Platinum**	["platinum"]
NCT0207	Pembrolizumab+**Pemetrexed**	["pemetrexed"]

Gold Answer: Every agent in Treatment Regimen **besides Pembrolizumab**. Eg, [Pemetrexed+ Carboplatin]

Figure 1: Sample exemplar hybrid reasoning question requiring semantic understanding of regimen structure: the model must recognize **Pembrolizumab** as the ICI named in the question and extract all other treatment agents from the visible regimen text as the held-out target field. For instance, in NCT0203, **Pembrolizumab** is the ICI and **Pemetrexed** is the additional agent used in the treatment regimen

1 Introduction

Clinicians and medical researchers often analyze tables of clinical trials with structured records where each row summarizes one study. Many questions over these tables can be answered by simple lookup: for example, what was the sample size of *Pembrolizumab drug trials*? But many questions require a deeper set of reasoning and cannot be answered immediately. For example, a reviewer studying Pembrolizumab drug trials which uses additional agents in the complete treatment regimen can surmise that, in a trial using Pembrolizumab + Pemetrexed, Pemetrexed medication is the additional agent in this treatment regimen. In these cases, the target value is not explicitly stored in a visible cell and cannot be obtained by simply selecting rows or projecting an existing column. Instead, it must typically be reasoned from row contents through normalization, classification, extraction, aggregation, or lightweight domain reasoning. We study this setting as *clinical trial reasoning*: given a question and a partially observed table, the clinicians must identify the relevant rows, ground the visible source field containing the evidence, and apply the correct row-wise transformation to produce

*These authors contributed equally.
†Corresponding authors.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 944–954
July 3-4, 2026 ©2026 Association for Computational Linguistics

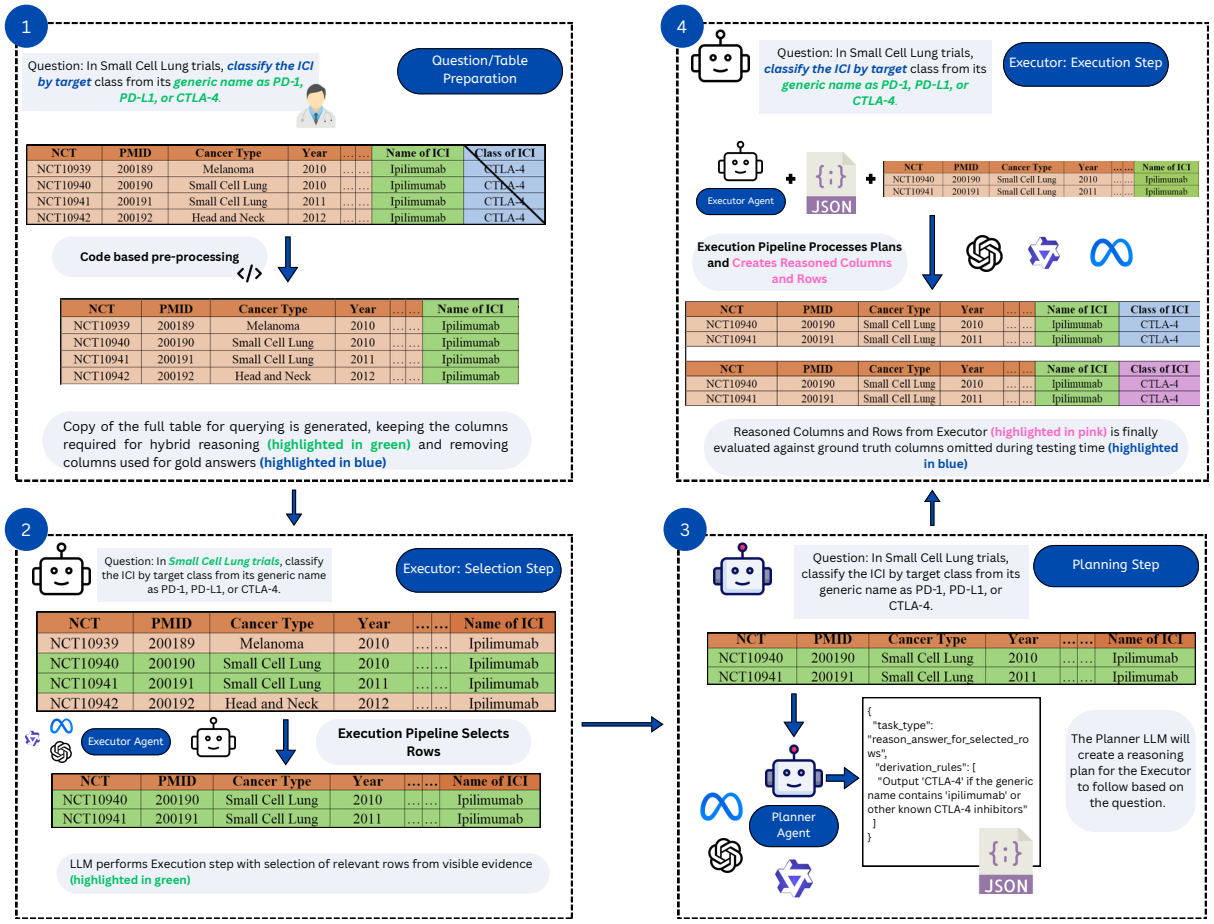


Figure 2: SCOPE for clinical tabular reasoning. Given a question and a copied visible table, SCOPE first prepares an inference-time table view by retaining the visible evidence columns and withholding the target column used for evaluation if there are any. The executor identifies the rows relevant to the question, the planner produces a structured reasoning plan over the selected table, and the executor follows this plan to generate the final row-aligned predictions. In this example, the method reasons over the *Name of ICI* column to recover the held-out *Class of ICI* field, and the resulting predictions are compared against the omitted ground-truth targets.

a target field that is often absent from the schema or from the evidence itself. This setting is related to table question answering and semantic parsing, but is only partially captured by existing table QA benchmarks and classical text-to-SQL formulations (Pasupat and Liang, 2015; Zhong et al., 2017; Yu et al., 2018; Chen et al., 2019; Nan et al., 2022).

Existing approaches do not fully address this problem. Text-to-SQL methods assume that the answer can be recovered through executable operations over explicit schema values, which is often not true when the target field is latent rather than stored (Zhong et al., 2017; Yu et al., 2018; Chowdhury et al., 2026). Prompting with large language models often under-structures the task, typically relying on a single generation step to retrieve rows, infer the evidence-bearing field, apply the transformation, and format row-aligned answers. More structured agentic and program-based approaches can improve reasoning, but can introduce additional execution overhead and brittle multi-step traces (Gao et al., 2023; Cheng et al., 2022; Yao et al., 2022; Glenn et al., 2024; Shi et al., 2024). So, the main research question we are trying to address is: *How effective is explicit hybrid planner-based decomposition in improving grounded row-level reasoning over partially observed clinical-trial tables, and do the resulting gains appear more clearly in row-level grounding than in aggregate answer recovery?*

To address these limitations, we introduce SCOPE, a planner-based framework for clinical querying that explicitly decomposes the task into row selection, structured planning, and execution. The executor grounds the question to relevant rows in the visible table and produces row-aligned outputs, while the planner interprets the question as

a structured transformation over the selected evidence. Concretely, the executor first identifies a candidate subset of rows, the planner converts the question into an explicit transformation specification, and the executor then applies that specification to generate aligned predictions. This decomposition makes the source-field choice, reasoning rules, and output constraints explicit prior to answer generation, improving interpretability and reducing ambiguity compared to direct prompting. By introducing an explicit planning interface between reasoning and execution, SCOPE better aligns the model behavior with the compositional structure of clinical trial querying.

We also construct a benchmark of 1,500 hybrid reasoning questions over oncology clinical-trial tables, where target attributes must be inferred from visible evidence rather than directly retrieved.

Unlike approaches that keep planning implicit in a single generation step or rely on heavier agentic pipelines, SCOPE introduces a lightweight planning interface that separates reasoning specification from execution while preserving grounding in the visible table. We evaluate SCOPE against zero-shot, few-shot, chain-of-thought, TableGPT2, BlendSQL, and EHRAgent, and find that explicit planning improves reasoning accuracy while offering a better accuracy-efficiency tradeoff than heavier agentic baselines.

Our main contributions are:

- We introduce SCOPE, a planner-based framework that decomposes clinical trial reasoning into row selection, source-field grounding, and structured execution.
- We formalize *clinical trial reasoning* as a distinct table understanding setting requiring inference over visible row evidence rather than direct cell retrieval.
- We construct a benchmark of 1,500 oncology clinical-trial questions and show that explicit planning outperforms prompting baselines while remaining more efficient than heavier agentic methods.

The code, scripts, and dataset are available at our project page.

Problem formulation. Given a question $\tilde{q}$ and a visible table $X = \{r_1, \ldots, r_m\}$ with a hidden target field, the goal is to recover a set of row-aligned predictions $\hat{G} = \{(r, \hat{y}_r)\}$ for the subset of rows relevant to the query. Each $\hat{y}_r$ must be derived from visible evidence in r rather than retrieved directly from a stored field. Errors may arise from incorrect row selection, incorrect source-field grounding, or incorrect transformation of visible evidence, making joint reasoning over these steps essential.

2 Related Work

We situate our work at the intersection of clinical datasets and table reasoning methods, focusing on settings where answers must be inferred rather than directly retrieved.

Related Datasets. Clinical datasets such as MIMIC III & IV (Johnson et al., 2016, 2023) and the eICU Collaborative Research Database (Pollard et al., 2018) have enabled large-scale modeling over structured and semi-structured medical records, while i2b2/n2c2 shared tasks (Uzuner et al., 2011) focus on extracting structured information from clinical text. In table reasoning, datasets such as TabFact (Chen et al., 2019) and FETAQA (Nan et al., 2022) study fact verification and free-form question answering over tables, assuming that answers are grounded in explicit table entries. Our benchmark construction is also motivated by recent domain-specific benchmarks that stress-test LLM capabilities on nuanced domain specific reasoning tasks (Choudhury et al., 2025, 2026). Given that and in contrast, our benchmark targets clinical-trial reasoning, where target attributes are often latent and must be inferred from visible row evidence through structured transformations.

Related Methods. Early table reasoning and semantic parsing approaches, such as Seq2SQL, TaBERT, and TAPAS (Zhong et al., 2017; Yin et al., 2020; Herzig et al., 2020), map natural language to executable queries or learned table representations and are strongest when the answer is explicitly represented in the schema. More recent LLM-based methods, including PAL and ReAct (Gao et al., 2023; Yao et al., 2022), introduce intermediate reasoning and action steps, while TableGPT2, BlendSQL, and EHRAgent (Li et al., 2024; Glenn et al., 2024; Shi et al., 2024) add stronger structure for reasoning over tables and clinical data. However, these methods either keep planning implicit within generation or rely on heavier agentic pipelines with iterative execution. In contrast, SCOPE explicitly decomposes the task into row selection, source-field grounding, and structured execution, making

the reasoning process more transparent and controllable while achieving a better accuracy-efficiency tradeoff.

3 Dataset

3.1 Dataset Construction

We begin with a seed set of 500 questions authored by a certified oncology researcher to reflect realistic evidence-review queries over a structured clinical-trial table containing bibliographic metadata, disease labels, treatment regimens, control-arm descriptions, endpoint text, trial phase, sample size, and follow-up information. Each seed example is paired with an executable SQLite query, which serves as a scaffold for generating new questions and aligned ground truth. We retain the subset of table-solvable questions that require hybrid reasoning rather than simple retrieval, and expand it to 1,500 questions through programmatic augmentation. For each retained pair (q, s), we generate a new pair $(\tilde{q}, \tilde{s}) = T(q, s)$ by applying a single atomic edit to the structured query, such as modifying the projection, simplifying predicates, swapping attribute values, relaxing thresholds, or changing the evidence field while preserving answerability. We keep only read-only variants that execute successfully and return non-empty results, then de-duplicate the resulting query patterns and rewrite each retained structured query into fluent natural language while preserving its exact semantics.

Ground-Truth Construction. For each final question $\tilde{q}$, the corresponding query $\tilde{s}$ identifies a set of relevant rows $R = \{r_1, \ldots, r_n\}$ together with a visible evidence field c_{src} used to derive the hidden target. We then apply a deterministic row-level reasoning function f to each row to produce the ground-truth target value $y_i = f(r_i[c_{\text{src}}])$ for every $r_i \in R$. This function implements the intended task semantics for normalization, categorization, boolean interpretation, and structured extraction, such as mapping an ICI mention to its class or brand name, determining endpoint role from endpoint text, bucketing follow-up duration or publication year, or extracting additional agents, schedules, regimen components, and control backbones from treatment text. To form the benchmark input, we expose only the visible table copy while omitting the target values $\{y_i\}_{i=1}^{n}$. The ground truth is stored separately as the row-aligned pairing $G = \{(r_i, y_i)\}_{i=1}^{n}$.

Clinical-Trial Table Statistic	Value
Rows	159
Columns	32
Unique trials (NCT)	105
Cancer types	19
ICI names	13

Table 1: Statistics of the underlying clinical-trial table used to construct the benchmark. The table covers trial-level oncology evidence, including study metadata, disease labels, treatment and control regimens, endpoint descriptions, and follow-up information, spanning 105 unique trials across 19 cancer types and 13 immune checkpoint inhibitor (ICI) names.

Crucially, each target value is derived from evidence already present in the table rather than authored independently or requiring external domain knowledge at evaluation time. This design ensures that the benchmark evaluates hybrid reasoning over visible evidence: models must infer latent attributes by interpreting existing table content, rather than generating plausible but unsupported answers.

3.2 Dataset Guidelines

To create the seed set, a certified clinician authored questions based on the IOTOX living evidence table[1]. They were asked to write realistic evidence-review questions while following three guidelines:

- Each question had to be grounded in visible table evidence.
- The expected answer had to be precise and structured, typically as a JSON value.
- The answer had to be inferable from the table itself, without requiring unsupported external knowledge.

3.3 Dataset Statistics

As shown in Tables 1 and 2, the benchmark is built from a structured clinical-trial table with 159 records and 32 columns, spanning 105 unique NCT identifiers, 131 unique PubMed IDs, 95 trial names, 19 cancer types, and 13 ICI names from 2010 to 2021. Although compact, the table is information-dense: many target outputs are not exposed as ready-made columns, but must be inferred from regimen text, endpoint descriptions, follow-up fields, and bibliographic metadata. At the benchmark level, the dataset contains 1,500 questions with

[1] https://iotox.living-evidence.com/

Question Statistic	Value
Total questions	1,500
Mean question length (tokens)	21.2
Target fields	31
Answer Types	**Value**
String	957
List	241
Boolean	224
Null-only	78

Table 2: Question and answer-level benchmark statistics. *Total questions* gives the size of the benchmark; *Mean question length* reports the average number of tokens per natural-language question; *Target fields* gives the number of distinct hidden target types covered by the dataset for reasoning . On the answer side, we report the number of questions whose gold supervision is primarily *string-valued*, *list-valued*, *boolean-valued*, or *null-only*, where null-only questions are those whose held-out target values are absent for all aligned rows.

an average length of 21.2 tokens and covers 31 distinct target fields. The answer space is also diverse, including string-valued, list-valued, boolean, and null-only outputs, highlighting that the benchmark evaluates hybrid reasoning across multiple target types rather than a single narrow form of table question answering.

4 The SCOPE System & Flow

Given a question $\tilde{q}$ and a visible table copy

$$X = \{r_1, \ldots, r_m\},$$

where the answer-bearing target field is omitted but row identifiers and visible evidence columns are preserved, SCOPE uses two LLM components: an *executor* and a *planner*. The executor grounds the question to relevant rows and produces the final row-level outputs, while the planner interprets the selected evidence as a structured reasoning problem. The planner may use the same backbone as the executor or a different one. We provide the row-selection, planning, and final-execution prompts in Appendix A.1.

The method has three steps. First, the executor identifies a candidate subset of relevant rows from the full visible table. Second, the planner receives the question and this candidate table and produces an explicit reasoning plan. Third, the executor follows that plan to generate the final row-aligned output table. In short, the planner determines *what* to derive and from *which* visible evidence, while the executor determines *where* that evidence occurs and returns the final answers.

Executor: Row Selection. The first step grounds the question $\tilde{q}$ to the visible table X. The executor selects a candidate subset of rows

$$\hat{R} = E_{\text{sel}}(\tilde{q}, X) \subseteq X.$$

At this stage, the executor is used only to identify relevant rows; it does not attempt to derive the final answer.

Planner: Structured Reasoning. The planner then receives the question $\tilde{q}$ and the candidate table $\hat{R}$ selected by the executor, and predicts a structured plan

$$\pi = P(\tilde{q}, \hat{R}) = \big(c_{\text{src}}, C_{\text{rel}}, \rho, o\big),$$

where c_{src} denotes the inferred visible source column that contains the supporting evidence, C_{rel} denotes the subset of visible columns retained for downstream execution, ρ denotes the reasoning rules needed to transform the source evidence into the target field, and o denotes the output constraints that specify the expected answer format.

Executor: Final Table Generation. Given the planner output π, the executor operates on the focused candidate table

$$X_\pi = \hat{R}[C_{\text{rel}}],$$

and produces the final row-aligned prediction table

$$\hat{G} = E_{\text{ans}}(\tilde{q}, \pi, X_\pi) = \{(r, \hat{y}_r) \mid r \in \hat{R}\}.$$

Each predicted value $\hat{y}_r$ is keyed by its row r, so the output remains aligned to the rows in the candidate table rather than collapsing the question into a single free-form answer.

Why Planning Helps. A key advantage of planning is that it explicitly separates decision points that are otherwise entangled in single-step generation. In particular, SCOPE decomposes (i) row grounding, (ii) source-field identification, and (iii) transformation into distinct stages, each of which can fail independently in direct prompting. By externalizing these decisions into a structured plan, the model avoids implicit assumptions about where evidence resides and how it should be transformed. This leads to more stable row alignment and reduces cascading errors compared to approaches where these steps are implicitly interleaved.

5 Experimental Setup

5.1 Baselines

We compare SCoPE against three baseline families under the same input setting. Every method receives the same question $\tilde{q}$ and visible table copy X, where the target field is omitted at inference time. When a method produces free text, SQL, or other intermediate outputs, we normalize its result programmatically into the shared row-aligned prediction format

$$\hat{G} = \{(\hat{r}_i, \hat{y}_i)\}_{i=1}^{k},$$

so that all systems are evaluated against the same held-out ground truth G.

The first family consists of zero-shot, few-shot, and chain-of-thought prompting (Brown et al., 2020; Wei et al., 2022), where the model is given the full visible table X and asked to recover the omitted target field directly without an explicit planning stage. The second includes BlendSQL (Glenn et al., 2024) and EHRAgent (Shi et al., 2024), which introduce more intermediate tabular structure reasoning through SQL-like compositional reasoning or heavier agentic execution. The third includes TableGPT2 (Li et al., 2024), a dedicated tabular reasoning model that jointly consumes the table and question to predict the hidden target field directly from the table representation.

5.2 Models and Hyperparameters

We evaluate three LLMs, chosen to span distinct model families, parameter scales, and post-training recipes while remaining fully reproducible: `Qwen3-30B-A3B-Instruct-2507` (Yang et al., 2025), `gpt-oss-20b` (Agarwal et al., 2025), and `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024). All three are open-weight, general-purpose instruction-tuned models that are not table-specialized, which lets us test whether SCoPE's gains generalize across model lineages and isolate the contribution of explicit planning from any table-specific pretraining. All generations use deterministic decoding with temperature 0.0 and top-p 1.0. Direct prompting baselines are allowed up to 2048 generated tokens per question. For SCoPE, the planner and executor are chosen from the same model pool, enabling both same-model and cross-model configurations. The row-selection step uses 768 tokens, the planner 1024, and the final executor step 2048, keeping execution lightweight while reserving more capacity for structured planning.

5.3 Evaluation Metrics

We evaluate each predicted table against the held-out ground-truth table using three grounded metrics after one-to-one row alignment. **Table F1** is our primary metric and measures how well a method recovers the correct hidden target values for the correct rows, balancing missed and spurious predictions. **Grounded Row Jaccard** gives a stricter overlap-based view by measuring the intersection-over-union between predicted and gold row-level outputs. **Grounded Fowlkes-Mallows** (Fowlkes and Mallows, 1983) complements these metrics by summarizing the balance between grounded precision and recall. Because the target field is omitted at inference time and restored only in held-out supervision, all three metrics evaluate recovery of hidden row-level attributes from visible evidence rather than direct copying.

Impact of row selection errors. Incorrect row selection directly affects all three metrics: missing relevant rows lowers recall, while selecting spurious rows lowers precision and overlap. As a result, the metrics jointly capture both row grounding and answer correctness.

6 Results and Discussion

6.1 Main Results

Across all backbone families, the most important result is that SCoPE consistently achieves the strongest grounded performance while remaining lightweight relative to heavier structured baselines. The framework is best overall on GPT-OSS and Qwen3, and tied on Table F1 while stronger on grounding metrics for Llama-3.3. The clearest gain appears for Qwen3, where explicit planning substantially improves row-aligned reasoning over direct prompting.

RQ1: Does explicit planning improve performance over direct prompting and tabular baselines? Yes. Table 3 shows that SCoPE is strongest or tied for strongest across all three model families on the grounded metrics. It consistently outperforms BlendSQL, EHRAgent, and TableGPT2, and also improves over the strongest prompting baselines in most settings. This supports the central claim of the paper: explicit planning is more effective than leaving row grounding and reasoning implicit inside a single generation step.

RQ2: Are these gains consistent across backbone models? Largely yes, although the size of

Method	Qwen3			Llama-3.3			GPT-OSS		
	F1 (%)	RJ (%)	FM (%)	F1 (%)	RJ (%)	FM (%)	F1 (%)	RJ (%)	FM (%)
Tabular Reasoning Methods									
BlendSQL	11.56	6.52	20.63	5.60	5.15	5.81	7.30	6.48	7.71
EHRAgent	32.99	29.79	33.74	30.99	28.07	31.69	34.85	31.23	35.65
Prompting Technique Methods									
Zero Shot	56.32	44.95	62.73	66.96	54.55	72.04	73.50	61.05	77.47
CoT	55.37	44.65	61.93	**70.87**	57.83	75.15	74.17	61.77	78.05
Few-Shot	54.74	44.09	61.48	69.38	56.56	74.05	73.99	61.55	77.85
Table Model Method									
TableGPT2	**F1:** 44.03			**RJ:** 33.78			**FM:** 50.81		
Ours: SCoPE									
SCoPE	**63.19**	**52.07**	**69.45**	**70.87**	**60.66**	**76.12**	**74.31**	**62.48**	**78.27**

Table 3: Per-model comparison across baselines and same-backbone SCoPE, where each listed model is used as *both* the planner and the executor. We report **F1** (Table F1), **RJ** (Row Jaccard), and **FM** (Fowlkes-Mallows). Higher is better for all metrics. Best values are shown in bold green. TableGPT2 is a table based model which does not have a Qwen3, Llama-3.3, or GPT-OSS backbone so it is tested individually.

the gain varies by model. For GPT-OSS, SCoPE improves over the strongest prompting baseline from 74.17 to 74.31 Table F1 while also yielding the best grounded overlap scores. For Llama-3.3, it matches the best prompting Table F1 (70.87) but improves Row Jaccard and Fowlkes-Mallows, indicating better row alignment. For Qwen3, the gains are largest, showing that the framework generalizes across backbones while being especially useful for weaker implicit reasoners.

RQ3: Which models benefit most from explicit planning? Qwen3 benefits the most. Its Table F1 increases from 56.32 under the best prompting baseline to 63.19 with SCoPE, with corresponding improvements in Row Jaccard and Fowlkes-Mallows. This suggests that explicit planning is particularly helpful when the underlying model is less reliable at performing row grounding and structured reasoning implicitly.

Note. In addition to Table F1, grounded Row Jaccard (RJ) complements evaluation by measuring intersection-over-union over row-aligned outputs, directly capturing joint correctness of row selection and target recovery. This is especially relevant here, where answers must be derived from the correct subset of rows rather than retrieved from explicit cells. Empirically, RJ shows stronger grounding quality: for Llama-3.3, SCoPE matches the best Table F1 (70.87) while increasing RJ from 57.83 to 60.66; for Qwen3, F1 improves by +6.87 (56.32 $\rightarrow$ 63.19) with a larger RJ gain of +7.12 (44.95 $\rightarrow$ 52.07); and for GPT-OSS, nearly unchanged F1 (74.17 $\rightarrow$ 74.31) still yields a consistent RJ increase (61.77 $\rightarrow$ 62.48).

Executor	Planner	F1(%)	RJ(%)	FM(%)
GPT-OSS	Qwen3	75.07	63.74	79.26
Qwen3	GPT-OSS	59.59	48.26	66.32
Qwen3	Llama-3.3	62.47	51.40	68.88
GPT-OSS	Llama-3.3	**75.12**	**63.88**	**79.28**
Llama-3.3	GPT-OSS	68.01	57.37	73.64
Llama-3.3	Qwen3	71.43	61.33	76.64

Table 4: Cross-model ablation of SCoPE. Rows report configurations where the executor and planner use different backbones. Higher is better. The best cross-model result in each metric is shown in bold green.

6.2 Cross-Model Ablation

We vary the planner and executor backbones to test whether planning and execution benefit from different model strengths. In particular, this ablation asks whether a stronger planner can improve performance even when the execution model is held fixed, and whether these two roles are better treated as complementary rather than identical.

Table 4 shows that changing the planner can indeed improve performance. The clearest example is GPT-OSS: when used as both planner and executor in Table 3, SCoPE reaches 74.31 Table F1, but replacing the planner with Llama-3.3 increases performance to 75.12, with similar gains in Row Jaccard and Fowlkes-Mallows. This suggests that the planner benefits from a stronger reasoning

Coder Model	F1(%)	RJ(%)	FM(%)
GPT-OSS	14.56	11.67	23.39
Qwen3	48.92	37.53	56.01
Llama-3.3	**63.40**	**51.55**	**69.10**

Table 5: Results for the planner-coder baseline across coder backbones. The same models are used for the planning and code generation execution stage. Best results are shown in bold green.

model even when execution remains fixed.

Another example appears in the Llama-3.3 executor family. In the same-backbone setting, Llama-3.3 reaches 70.87 Table F1, but replacing the planner with Qwen3 improves performance to 71.43, with corresponding gains in Row Jaccard and Fowlkes-Mallows. Together with the GPT-OSS results, this suggests that changing the planner can improve performance even when the executor is held fixed.

6.3 Planner-Coder Baseline

We include the planner-coder setting as an ablation of whether execution should be made more explicit through programmatic synthesis. The idea is that, if the planner already produces a structured reasoning trace, then compiling that trace into deterministic Python code could make execution more transparent and reproducible. This tests an alternative to grounded LLM execution: executing the plan by first translating it into code.

Table 5 shows that this formulation performs substantially worse than SCOPE across all backbones. The drop is especially large for GPT-OSS (from 74.21 to 14.56 Table F1) and Qwen3 (from 63.19 to 48.92), while even the best planner-coder result with Llama-3.3 (63.40) remains below the strongest same-backbone SCOPE systems. This suggests that code-generation-based execution is more brittle: the model must infer the procedure, translate it into executable code, and ensure that the code still matches the table context and output format. Overall, hybrid clinical table reasoning appears to benefit more from constrained grounded execution than from open-ended code synthesis.

7 Model Cost Effectiveness Analysis

Figure 3 compares answer quality against token budget in the Qwen-based setting. Although TableGPT2 is not Qwen-based, we include it as a reference point for cost-effectiveness against a dedicated table model. SCOPE lies on the strongest

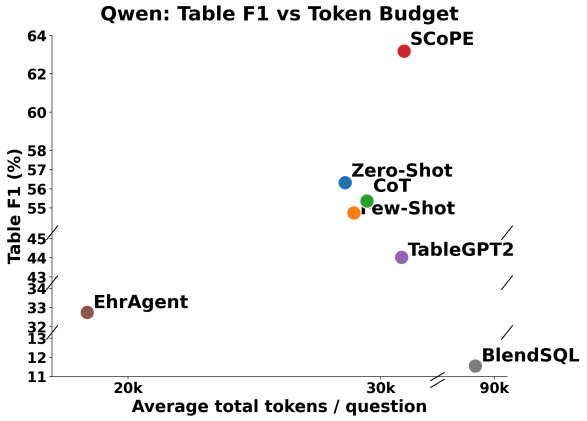


Figure 3: Cost-effectiveness comparison for Qwen-based methods. The x-axis shows average total tokens per question and the y-axis shows **Table F1**.

accuracy-cost frontier, achieving the highest Table F1 at roughly 63% with only a modestly larger token budget than direct prompting baselines. Zero-shot, few-shot, and chain-of-thought prompting cluster in the 28k-30k token range but remain in the mid-50% F1 range, while EHRAgent is cheaper but much less accurate and BlendSQL is both the most expensive and the weakest. Overall, SCOPE adds limited planning overhead while yielding the best accuracy, making it the most favorable cost-effective choice in this comparison.

8 Conclusion

We introduced SCOPE, a planner-based framework for clinical trial table reasoning that separates row grounding, structured planning, & row-level execution over visible evidence. Across same-backbone & cross-model settings, explicit planning consistently improves grounded reasoning performance over direct prompting & stronger structured baselines such as BlendSQL, EHRAgent, & TableGPT2. The gains are especially pronounced for weaker implicit reasoners while remaining competitive for stronger backbones. These findings position clinical trial reasoning as a distinct table understanding problem & suggest that lightweight planner-executor decomposition is an effective and cost-conscious approach for recovering latent clinical attributes from partially observed oncology tables.

Limitations

Our study has four main limitations. First, we evaluate strong open-weight/open-access models but do not include the latest frontier proprietary systems, so the measured gains from explicit planning may

change under stronger closed models. Second, we analyze same-backbone, cross-model, and planner-coder settings, but we do not fine-tune planner and executor roles separately; role-specialized training could alter the observed tradeoffs. Third, we do not run a dedicated memorization analysis. Because the benchmark is derived from a compact table and programmatically expanded from a seed set, some patterns may be partially recoverable from prior exposure rather than purely from visible evidence. Fourth, the benchmark is limited to a single oncology clinical-trial table, so broader validation across additional clinical domains, schemas, and larger real-world evidence tables is still needed.

Ethics Statement

This work supports oncology evidence review and clinical-trial exploration, not patient-specific diagnosis or treatment decisions. SCOPE uses public trial-level records only and does not use patient notes, PHI, or PII. No human subjects were recruited, and no new patient data were collected or released.

Because LLM outputs can be incorrect or overconfident, we emphasize auditability and grounding in visible table evidence. Evaluation is based on row-aligned recovery of omitted fields, and the pipeline exposes planning and execution steps for inspection. We also restrict execution to read-only operations and apply schema validation and timeout safeguards. Our benchmark construction keeps only valid read-only variants, which improves consistency but may bias coverage toward frequent query patterns.

While the seed questions are authored by verified clinicians, he benchmark reflects a constrained setting: a single oncology clinical-trial table, without longitudinal patient records or free-text clinical narratives.

Performance on this benchmark should not be interpreted as clinical reliability. Outputs from SCOPE require expert review before practical use. Real-world deployment should include access controls, logging, and clear warnings about model limitations and possible hallucinations.

Acknowledgment

This research was supported by the Mayo Clinic and Arizona State University Alliance for Health Care Collaborative Research Seed Grant Program (Award ID: AWD00041508; Sponsor Award ID: ARI-358187) for the project: 'Artificial intelligence-assisted digital, living, interactive clinical practice guidelines for cancer providers and patients

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, and 1 others. 2022. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*.

Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand, and Vivek Gupta. 2026. Better call CLAUSE: A discrepancy benchmark for auditing LLMs legal reasoning capabilities. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5776–5818, Rabat, Morocco. Association for Computational Linguistics.

Manan Roy Choudhury, Anirudh Iyengar Kaniyar Narayana Iyengar, Shikhhar Siingh, Sugeeth Puranam, and Vivek Gupta. 2025. TABARD: A novel benchmark for tabular anomaly analysis, reasoning and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21783–21817, Suzhou, China. Association for Computational Linguistics.

Suparno Roy Chowdhury, Tejas Anvekar, Manan Roy Choudhury, Muhammad Ali Khan, Kaneez Zahra Rubab Khakwani, Mohamad Bassam Sonbol, Irbaz Bin Riaz, and Vivek Gupta. 2026. Fd-nl2sql: Feedback-driven clinical nl2sql that improves with use. *arXiv preprint arXiv:2604.15646*.

Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language

models. In *International conference on machine learning*, pages 10764–10799. PMLR.

Parker Glenn, Parag Dakle, Liang Wang, and Preethi Raghavan. 2024. Blendsql: A scalable dialect for unifying hybrid question answering in relational algebra. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 453–466.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4320–4333.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. Table-gpt: Table fine-tuned gpt for diverse table tasks. *Proceedings of the ACM on Management of Data*, 2(3):1–28.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, and 1 others. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8413–8426.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3911–3921.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

A Supplementary Material

A.1 Planner Prompts

We provide the three inference-time prompt variants used by SCOPE below. Together, they implement the core stages of the pipeline: the selector-executor prompt identifies the candidate rows relevant to the question, the planner prompt converts the question and candidate table into an explicit reasoning plan, and the planner-executor prompt uses that plan to produce the final row-aligned predictions. Including these prompts makes clear how SCOPE externalizes row grounding, source-field selection, and answer generation rather than leaving them implicit in a single generation step.

More information on error type analsysi and prompts utilized for the baseline The code are available at our project code base underneath the *appendices* tab.

Prompt A: Plan Executor

```
You are the final executor stage of a clinical-trials
    ↪ table reasoning system.
Use the planner plan and the visible candidate table to
    ↪ derive the answer for each row.

Return ONLY valid JSON in this exact shape:
{"predictions":[{"table_row_id": 1, "answer": ...}]}

Rules:
- Return one prediction for each row in the candidate
    ↪ table.
- Use the integer "rowid" values from the candidate table
    ↪ as table row id.
- Use the planner's inferred source column and derivation
    ↪ rules when helpful.
- Derive the answer from visible values and visible row
    ↪ context.
- Do not assume the answer already appears verbatim in the
    ↪  table.
- Follow the planner's normalization rules whenever
    ↪ possible.
- Do not add explanations outside the JSON.
```

Prompt B: Planner

```
You are the planner stage of a clinical-trials hybrid
    ↪ querying system.

Overall context:
{{overall_context}}

Your job is to understand the essence of the question and
    ↪ help another model derive the answer implied by
    ↪ the question for each selected row.
The answer may need classification, normalization,
    ↪ extraction, boolean inference, or simple
    ↪ transformation.
Focus on what must be derived from visible source values
    ↪ and row context.

Return ONLY valid JSON in this exact shape:
{
  "task_type": "derive_answer_for_selected_rows",
  "inferred_source_column": "...",
  "derived_field_description": "...",
  "answer_type": "string|boolean|number|list|object",
  "relevant_columns": ["__rowid__", "..."],
  "row_filter_restatement": "...",
  "derivation_rules": ["...", "..."],
  "normalization_rules": ["...", "..."],
  "output_constraints": ["...", "..."],
  "notes": "..."
}

Guidance:
- 'inferred_source_column' should name the visible column
    ↪ that most likely contains the source evidence the
    ↪  executor should derive from.
- 'relevant_columns' should be a focused subset of the
    ↪ visible headers that the executor needs in the
    ↪ final step.
- 'derivation_rules' should explain how to derive the
    ↪ answer from visible source values and context.
- 'normalization_rules' should specify canonical output
    ↪ formatting when relevant.

Question:
{{question}}

Visible headers:
{{headers_json}}

Candidate row count:
{{candidate_row_count}}

Candidate preview JSON:
{{candidate_preview_json}}
```

Prompt C: Selector-Executor

```
You are the executor stage of a clinical-trials table
    ↪ reasoning system.
Your current job is only to identify which visible rows
    ↪ satisfy the question.
Do not derive the final answer yet.

Return ONLY valid JSON in this exact shape:
{"selected_row_ids":[1,2,3],"selection_reason":"...","
    ↪ needs_derivation":true}

Rules:
- Use only the visible table content.
- Use the integer "__rowid__" values from the table.
- If no rows satisfy the question, return {"
    ↪ selected_row_ids":[],"selection_reason":"...","
    ↪ needs_derivation":true}.
- Do not add explanations outside the JSON.

Question: {{question}}

Visible headers:
{{headers_json}}

Visible table (CSV):
{{table_csv_text}}
```

Expert-Guided Schema-Based Structured Extraction from CONSORT Diagrams Using Vision-Language Models

Damian Stachura, Bartosz Przechera, Monika Opałek,
Ewelina Sadowska, Ewa Borowiack, and Artur Nowak
Evidence Prime, Krakow, Poland
damian.stachura@evidenceprime.com

Abstract

Visual-language models (VLMs) are rapidly advancing on tasks that require visual understanding of text, tables, plots, and diagrams. Yet extracting structured information from text-heavy scientific diagrams remains challenging, as it requires not only OCR but also recovery of layout, grouping, and flow relationships. We study this problem in the context of CONSORT flow diagrams, which summarize participant screening, randomization, follow-up, and analysis in randomized controlled trials. We introduce a 200-example benchmark of PubMed Central diagrams, annotated by a biomedical team specializing in systematic literature reviews and clinical evidence extraction, and evaluate schema-constrained CONSORT extraction across proprietary and open-weight model families. Using structure-aware metrics, we compare single-pass and stepwise extraction strategies. Expert-guided single-pass extraction performs best for proprietary frontier models, with Gemini 3 Pro achieving the strongest overall results, whereas stepwise prompting improves less capable open-weight models on challenging arm-level extraction. These results offer practical deployment guidance and suggest that high-quality schema-constrained extraction is feasible, but not yet solved.

1 Introduction

Scientific articles contain rich structured knowledge, but extracting it remains challenging because evidence is distributed across multiple modalities, including narrative text, tables, charts, and diagrams. Recent vision-language models (VLMs) and multimodal document understanding systems have improved the ability to process visually rich scientific documents, making it increasingly feasible to extract information directly from figures and other non-textual layouts (Hu et al., 2023; Fujitake, 2024; Pramanick et al., 2024; Wang et al., 2025). However, text-heavy visual artifacts remain difficult, as successful extraction requires not only reading embedded text but also understanding layout, grouping, topology, and numerical relationships between visual elements.

A particularly important case arises in randomized controlled trials (RCTs), where key study information is often presented visually. Among these, CONSORT flow diagrams are especially valuable because they summarize participant screening, exclusions, randomization, follow-up, and analysis populations (Moher et al., 2010; Hopewell et al., 2011, 2025). However, these diagrams are highly heterogeneous in wording, layout, and completeness, making structured extraction of patient-flow information a challenging multimodal problem. An example is shown in Figure 1.

Prior work on CONSORT has largely focused on reporting support rather than figure understanding, including diagram generation, checklist extraction, and reporting-quality assessment (O'Leary et al., 2019; Wang et al., 2020; Laï-king and Paroubek, 2025). While broader research has explored methodological and multimodal evidence extraction from RCT publications (Hoang et al., 2023; Ghosh et al., 2024; Ji et al., 2026), extracting structured patient-flow data directly from CONSORT diagrams remains underexplored.

To address this gap, we introduce a benchmark for structured extraction from CONSORT flow diagrams. We collaborate with biomedical experts to define a schema capturing key patient-flow information and annotate a dataset of diagrams with substantial visual and reporting variability. Using this benchmark, we evaluate multiple VLM-based extraction strategies, including end-to-end schema filling with different prompting styles and stepwise approaches that decompose the task into smaller, visually coherent subproblems.

Our contributions are as follows:

- We introduce a 200-example benchmark for

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 955–969
July 3-4, 2026 ©2026 Association for Computational Linguistics

structured extraction from CONSORT flow diagrams, with schema-aligned annotations prepared in collaboration with biomedical experts.

- We define and report a structure-aware evaluation suite for nested CONSORT flow outputs, including group-level, leaf-level, strict field-level, and diagnostic error metrics.
- We provide a capability profile across proprietary and open-weight VLMs under single-pass and stepwise prompting, and derive practical model-regime guidance for deployment.

2 Related work

Prior automation around CONSORT has primarily focused on reporting support rather than figure understanding. For example, O'Leary et al. (2019) proposed automatic generation of CONSORT diagrams from clinical-trial databases, while Wang et al. (2020) introduced CONSORT-NLP for generating reporting checklists from article PDFs. More recent work has evaluated large language models for CONSORT-oriented reporting-quality assessment (Laï-king and Paroubek, 2025). Related research has explored information extraction from RCT publications more broadly, including methodological extraction (Hoang et al., 2023), PICO extraction (Ghosh et al., 2024), and structured evidence extraction pipelines such as TrialMind and LatteReview (Wang et al., 2024b; Rouzrokh et al., 2025). While these approaches demonstrate progress toward structured clinical evidence extraction, they do not address recovery of patient-flow information from CONSORT diagrams, where key evidence is expressed visually.

The closest technical work comes from diagram and visually rich document understanding. Prior research has highlighted the challenges of OCR and figure-text extraction in scientific documents (Kim and Yu, 2011), while more recent approaches combine OCR, element detection, and generative reasoning for diagram parsing (Arbaz et al., 2024; Deka and Devereux, 2025). Multimodal systems have further explored diagram understanding (Hu et al., 2023), layout-aware document processing (Fujitake, 2024), and schema-constrained extraction (Wang et al., 2024a). These studies suggest that extraction from text-heavy visuals benefits from explicit structure recovery rather than treating images as generic captioning targets.

A related line of work studies reasoning over diagrams and flowcharts. Benchmarks such as FlowchartQA and FlowVQA evaluate multimodal reasoning in flowchart settings (Tannert et al., 2023; Singh et al., 2024), while FlowLearn highlights persistent weaknesses of VLMs in OCR, counting, and structural reasoning (Pan et al., 2024). On the modeling side, approaches such as TextFlow and arrow-guided VLMs explicitly incorporate graph structure and directional relationships to improve flowchart understanding (Ye et al., 2024; Omasa et al., 2025). These findings are particularly relevant because CONSORT diagrams represent participant flow through structured, text-bearing graph elements.

Recent work also explores agentic and tool-augmented approaches to multimodal extraction. For example, Chen et al. (2025) propose a multi-agent system for extracting structured information from scientific graphics, while systems such as ViperGPT demonstrate how visual reasoning can be decomposed into executable tool calls (Surís et al., 2023). These approaches suggest that modular reasoning and explicit structure recovery may be beneficial for complex visual extraction tasks.

Despite these advances, no prior work specifically targets structured extraction of patient-flow data from published CONSORT diagrams, leaving a clear gap for dedicated benchmarks and systematic evaluation of extraction strategies.

3 Benchmark

To evaluate extraction performance, we constructed a dedicated benchmark for CONSORT flowchart data extraction. This task remains relatively underexplored, and publicly available benchmarks are scarce, which motivated us to curate our own evaluation dataset.

3.1 Data Acquisition

We collected candidate CONSORT flow diagrams from PubMed Central articles published under CC-BY 4.0 licenses, which permit figure reuse. We first assembled a pool of 500 diagrams by randomly sampling eligible literature. From this pool, we selected 180 examples for benchmark release through a focused curation pass designed to preserve visual diversity, including layout styles, branch depth, and labeling conventions, as well as difficulty diversity, ranging from simple linear flows to structurally complex cases such as nested branches and partial

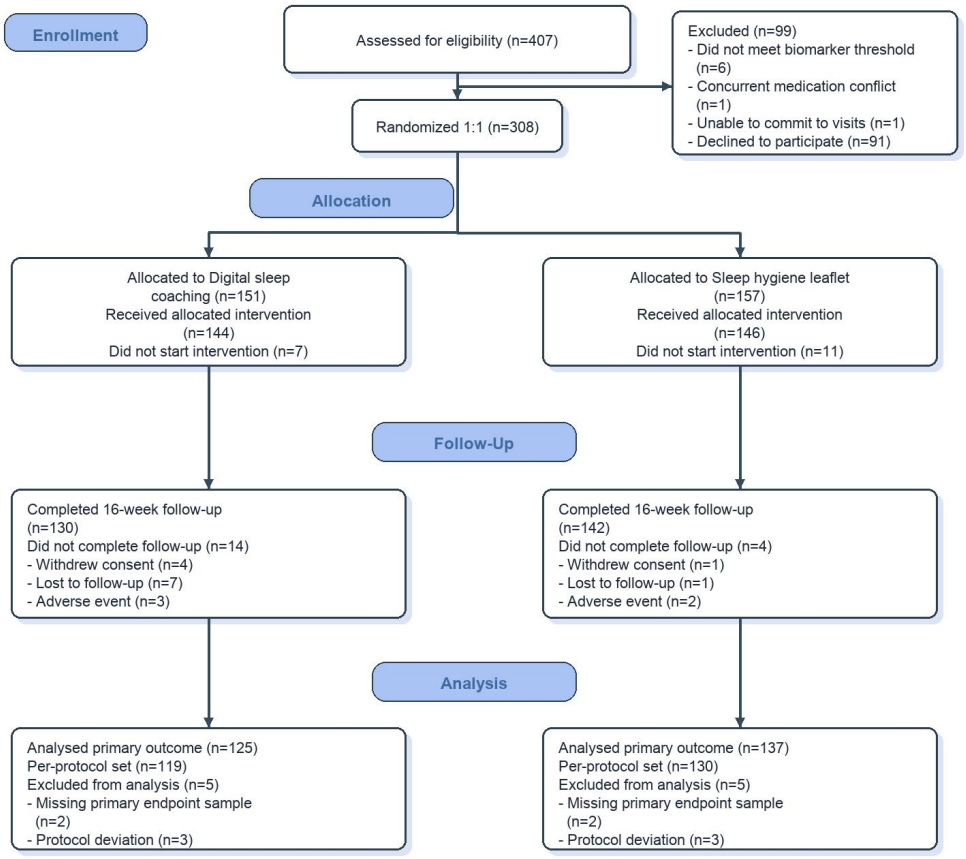


Figure 1: Synthetic CONSORT flowchart showing participant screening, randomization, follow-up, and analysis across intervention and control arms.

assessments. We focused on selecting only literature reporting RCTs. Additionally, we added 20 expert-curated examples that were considered challenging to reach a quick consensus on regarding how the data should be extracted. These 200 examples constitute the full evaluation benchmark used in this paper.

The 20 expert-curated diagrams are included as a deliberate stress-test subset to increase structural difficulty and probe failure modes under harder conditions. All systems are evaluated with the same fixed benchmark, schema, and scoring pipeline, so comparisons remain transparent and directly comparable across models and prompting regimes. We therefore interpret the reported results as a capability stress profile for structured CONSORT extraction rather than as a prevalence estimate for the full population of published CONSORT diagrams.

Each sample consists of a CONSORT flowchart paired with a structured representation following our schema. Annotation and verification were carried out by a team of three biomedical experts with background from masters and doctoral-level studies in biomedical domains and prior industry experience, specialized in systematic literature reviews and clinical evidence synthesis. The team aligned study-level and arm-level counts, exclusions, and analysis populations to canonical schema fields despite substantial variation in wording and layout.

3.2 Data Structure

CONSORT flowcharts are highly heterogeneous in both visual layout and textual content. Authors vary in how they organize diagrams, which counts they report, and the level of detail included in individual nodes. Despite this variability, most flowcharts can be mapped to a small number of recurring semantic components.

A typical CONSORT flowchart includes the following components:

- **Enrollment and screening:** the number of individuals assessed for eligibility, along with the number excluded prior to randomization and the reported reasons for exclusion.
- **Randomization and allocation:** the number of participants randomized and assigned to each study arm.
- **Follow-up and intervention status:** for each study arm, the number of participants who received the allocated intervention, did not receive it, completed treatment, discontinued treatment, or were lost to follow-up, together with corresponding reasons when reported.
- **Analysis:** the number of participants included in the primary outcome analysis for each arm, along with any exclusions from analysis and their reported reasons.

Our schema captures this structure at two levels. At the study level, it represents overall eligibility, pre-randomization exclusions, randomization, and the set of study arms. At the arm level, it captures participant trajectories within each arm, including allocation, treatment receipt, discontinuation, follow-up, and analysis. Supporting objects store finer-grained breakdowns of ineligibility, intervention discontinuation, analysis exclusions, and partial assessments. Both levels are illustrated in Figure 2 and summarized compactly in Table 1.

3.3 CONSORT Schema Fields

The target schema for CONSORT flow extraction is organized around a study-level object and a collection of per-arm objects. The fields we included in our schema were based on the CONSORT diagram schema presented in the literature (Hopewell et al., 2025) and we have chosen the fields whose experts flagged as most important in the process of random controlled trial analysis.

The study-level object tracks eligibility, exclusions, randomization, washout structure, and arm containers. Each arm object tracks allocation, treatment receipt and completion, follow-up losses, analysis populations, exclusions, and optional nested branches. A detailed field-by-field glossary for both levels is provided in Appendix A.

3.4 Data Annotation Protocol

The dataset was annotated using a two-stage procedure:

Component	Main content
Study-level flow	Eligibility, pre-randomization exclusions, randomization totals, washout indicator, arm containers, optional study-level analysis populations.
Arm identity	Arm label and ordering used for branch alignment.
Arm treatment flow	Allocation, treatment receipt/non-receipt, completion/non-completion, discontinuation reasons.
Arm follow-up and analysis	Follow-up losses, analysis populations, analysis-stage exclusions.
Optional advanced structure	Partial assessments and nested randomization branches.

Table 1: Compact summary of the CONSORT extraction schema used in the main experiments.

- Each expert independently mapped data from the CONSORT diagram to our predefined schema, which had been developed with the support of the same group of experts.
- Disagreements were adjudicated by the expert panel, and corrected labels were rechecked before dataset freeze.

4 Methods

Using this benchmark, we evaluated three VLM-based strategies for structured extraction from CONSORT flow diagrams in a zero-shot setting. The comparison was designed to isolate the effects of prompt scope, task decomposition, and post hoc verification while keeping the target output schema fixed.

- **Direct end-to-end extraction with basic prompt:** a single-pass baseline in which one VLM call using a basic system prompt receives the full diagram and optional caption and returns the complete target structure.
- **Direct end-to-end extraction with expert-guided prompt:** a single-pass baseline in which one VLM call with a prompt crafted by domain experts receives the full diagram and optional caption and returns the complete target structure.
- **Stepwise extraction:** a multi-pass strategy that reuses the same full diagram across a sequence of narrower prompts for study-level fields, washout detection, arm inventory, and per-arm flow extraction.

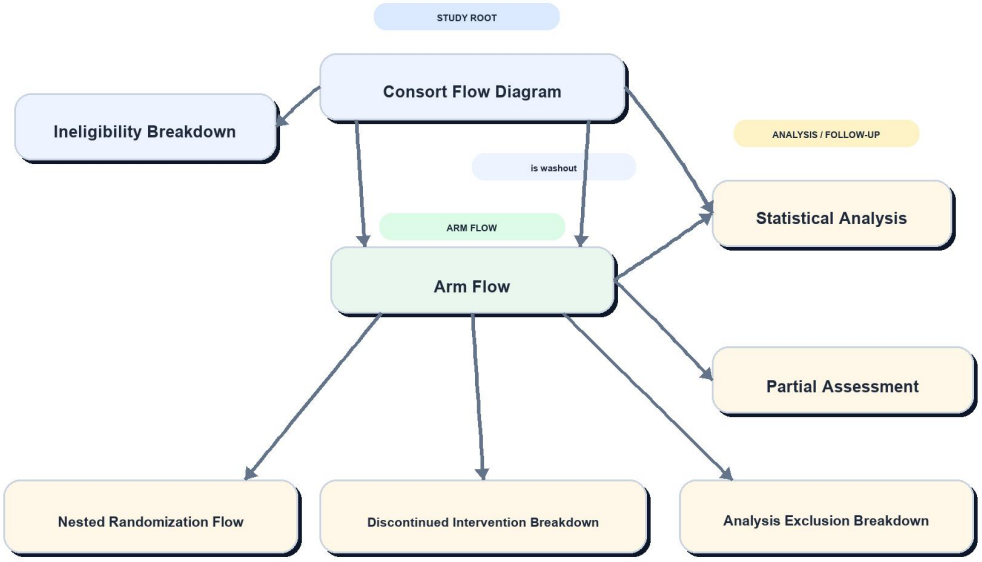


Figure 2: High-level schema of the CONSORT flow data model. The study-level CONSORT Flow Diagram captures overall enrollment and ineligibility information, a washout indicator, and arrays of Arm Flow objects before and after washout. Each study arm then records its own participant flow, including nested randomization, discontinuation, partial assessments, statistical analysis, and analysis-exclusion components.

4.1 Direct End-to-End Extraction

The first approach is the single pass setting. In this setting, a single VLM call is provided with the full CONSORT diagram, optionally accompanied by its caption, and is tasked with populating the complete CONSORT diagram structure in one step. This formulation serves as a direct benchmark of diagram understanding, as the model must jointly infer study-level counts, arm-level participant flows, and analysis populations without intermediate decomposition.

We evaluate two variants of prompting within this setting. The first uses a minimal prompt that guides the model to extract information according to the target schema, without additional domain-specific guidance. The second employs an extended prompt enriched with detailed definitions of schema fields and curated synonym lists, prepared by domain experts, to better align the model's interpretation with CONSORT reporting conventions.

4.2 Stepwise Extraction

The second approach is the stepwise setting. Instead of asking for the whole structure at once, the extractor runs a sequence of scoped passes over the same full diagram. The first pass extracts study-level enrollment and randomization fields, the second detects whether an explicit washout phase is present, and the third identifies the top-level arm inventory. The extractor then runs one additional pass for each arm to recover the detailed arm flow. These partial outputs are compiled deterministically into the final diagram structure. Importantly, this strategy does not crop the image into local regions. Rather, it keeps the full diagram available at every step and narrows the prompt scope to reduce structural ambiguity.

5 Experiments

This section presents the experimental setup and evaluation results for CONSORT diagram extraction.

The goal of the experiments is to characterize extraction capability across model families and scales, not only to rank a single best system. We therefore report structure-aware metrics, strict exact-match, and subgroup diagnostics to expose residual failure modes even when top-line scores are high.

5.1 Evaluation Metrics

We evaluate CONSORT diagram extraction using a structure-aware metric suite designed for nested participant-flow outputs. Since the task requires both accurate value extraction and correct assign-

ment of values to the appropriate arm, phase, or nested branch, our evaluation emphasizes structural correctness in addition to leaf-level accuracy.

Before scoring, both expected values and predicted outputs are normalized using the same schema semantics as the extraction system. This includes canonicalizing the CONSORT diagram representation, normalizing strings via case folding and whitespace collapsing, preserving exact integer counts, and resolving schema-level conventions such as washout structure. Object-valued lists are aligned by semantic identity rather than raw position, and objects representing arms, partial assessments, nested randomizations, exclusion reasons, and analysis populations are matched by their names.

We focus on three metrics that capture overall quality, structural correctness, and arithmetic coherence of extracted CONSORT flow representations.

CONSORT Group F1. This is the main structure-aware metric. It summarizes correctness across grouped components of the CONSORT diagram, including study-level structure, top-level arms, washout branches, nested randomizations, reason groups, partial assessments, and analysis populations. This metric is particularly important because a model may extract correct counts but still attach them to the wrong arm or branch.

Leaf Micro F1. This metric averages correctness over aligned atomic leaves. Numeric leaves are scored by exact match, while free-text identity fields such as arm names, assessment names, and reasons are evaluated using normalized string similarity. It reflects extraction quality at the value level once the hierarchical structure has been aligned.

Exact Field Accuracy. This metric measures the fraction of aligned schema fields that match exactly after normalization. It is stricter than the F1-style metrics because it does not award partial credit at the field level, but unlike full-example exact match it still captures partial success within a diagram.

Dataset-Level Aggregation. Metrics are computed per example and averaged over the evaluation set. We also report a normalized exact-match rate as a supplementary indicator, but not as a primary metric due to its stricter nature compared to the structure-aware measures above.

5.2 Technical Details

For the experiments, we evaluated all systems on the same 200-example benchmark. All runs used deterministic decoding with temperature 0 in a zero-shot setting. We did not provide figure captions during benchmarking, so the reported results reflect extraction from the diagram alone. For the stepwise system, the full image remained available at every stage.

Our primary model-selection objective was to measure the practical upper bound of capability under a fixed extraction protocol. We therefore used proprietary frontier models across size tiers to test what the strongest current systems can do on this task. The proprietary ablation in Table 2 shows a consistent pattern that expert-guided single-pass prompting is the strongest setting for these models.

We then evaluated open-weight models to estimate how far smaller openly available VLMs remain from proprietary systems on the same benchmark. This comparison is capability profiling under a common schema and evaluation pipeline, not a parameter-matched fairness claim. We also test smaller open-weight models because CONSORT extraction can be used as an agentic subtask in larger evidence-extraction workflows, where efficient local models may offer practical advantages.

Prompting setups are intentionally reported asymmetrically across regimes. For proprietary models, we run full basic/expert/stepwise ablations to establish the best prompting policy. For the newer open-weight 30B and Gemma runs, we focus on expert single-pass versus stepwise because this is the decision-relevant comparison after the proprietary ablation.

To assess the range of systems that can address this task, we used:

- GPT-5.1 and GPT-5 (OpenAI, 2025) mini with low reasoning effort deployed via Azure AI,
- Gemini 3 Flash Preview and Gemini 3 Pro Preview (DeepMind, 2025) with low thinking level accessed through the Google API,
- Qwen3-VL 8B Instruct (Bai et al., 2025) with 8-bit quantization (Ollama[1]), run locally on a MacBook M4 Max,
- Qwen3-VL 30B A3B Instruct (Bai et al., 2025) and Gemma-4-26B-A4B-it (Google

[1] https://ollama.com/

System	Mode	Group F1	Leaf Micro F1	Exact Field Acc.	Exact Match
Gemini 3 Pro	basic prompt	0.903	0.896	0.822	0.210
Gemini 3 Pro	expert prompt	**0.976**	**0.971**	**0.944**	**0.705**
Gemini 3 Pro	stepwise	0.903	0.891	0.809	0.155
Gemini 3 Flash	basic prompt	0.905	0.896	0.817	0.210
Gemini 3 Flash	expert prompt	0.964	0.957	0.926	0.625
Gemini 3 Flash	stepwise	0.907	0.895	0.818	0.165
GPT-5.1	basic prompt	0.874	0.843	0.743	0.050
GPT-5.1	expert prompt	0.872	0.832	0.732	0.065
GPT-5.1	stepwise	0.819	0.800	0.654	0.000
GPT-5 mini	basic prompt	0.850	0.812	0.675	0.055
GPT-5 mini	expert prompt	0.868	0.824	0.718	0.065
GPT-5 mini	stepwise	0.820	0.789	0.636	0.000

Table 2: Proprietary-model prompting ablation on the 200-example CONSORT benchmark (zero-shot, deterministic decoding, diagram-only input without figure captions). Expert-guided single-pass prompting is the strongest proprietary setting and is therefore used as the reference for open-weight comparisons.

Open-weight system	Mode	Group F1	Leaf Micro F1	Exact Field Acc.	Exact Match
Qwen3-VL 8B Instruct	expert prompt	0.701	0.593	0.375	0.000
Qwen3-VL 8B Instruct	stepwise	0.571	0.539	0.063	0.005
Qwen3-VL 30B A3B Instruct	expert prompt	0.601	0.495	0.379	0.000
Qwen3-VL 30B A3B Instruct	stepwise	0.633	0.611	0.435	0.005
Gemma-4-26B-A4B-it	expert prompt	0.635	0.541	0.068	0.000
Gemma-4-26B-A4B-it	stepwise	0.622	0.564	0.407	0.000

Table 3: Open-weight model comparison on the same 200-example benchmark and evaluation pipeline (zero-shot, deterministic decoding, diagram-only input).

DeepMind, 2026) via OpenRouter[2].

5.3 Results

Table 2 establishes prompting behavior on proprietary models. Across Gemini variants, expert-guided single-pass prompting clearly outperforms both basic and stepwise prompting on all primary metrics. Gemini 3 Pro with expert prompting is the strongest proprietary configuration overall, reaching 0.976 group F1, 0.971 leaf micro F1, 0.944 exact field accuracy, and 0.705 exact match.

Table 3 reports open-weight results for Qwen3-VL 8B Instruct, Qwen3-VL 30B A3B Instruct, and Gemma-4-26B-A4B-it in expert single-pass and stepwise modes. In this lower-capability regime, expert single-pass fails on complex arm-level extraction for Qwen3-VL 30B A3B Instruct and Gemma-4-26B-A4B-it. For Qwen3-VL 30B, stepwise improves Group F1 from 0.601 to 0.633, Leaf Micro F1 from 0.495 to 0.611, and Exact Field Accuracy from 0.379 to 0.435; for Gemma-4-26B-A4B-it, stepwise improves Leaf Micro F1 from 0.541 to 0.564 and Exact Field Accuracy from 0.068 to 0.407 (with a small Group F1 decrease from 0.635 to 0.622).

This suggests that decomposition can help smaller models, but the current full-context stepwise design is still limited. A promising next step is arm-localized extraction: first detect arms, then run extraction on per-arm crops instead of repeatedly presenting the full diagram context.

From a deployment perspective, these results support a regime-dependent policy: use expert single-pass prompting for frontier proprietary systems, and prefer stepwise prompting for smaller open-weight systems when arm-level complexity is high.

Despite strong top-line scores for the best systems, strict exact match remains below one for all completed models (best: 0.705), and the appendix subgroup diagnostics show persistent errors on structurally complex components. This supports the benchmark's utility as a non-trivial capability test rather than an already-solved task.

5.4 Error Analysis

We observe several recurring error types in our results. First, some errors arise from ambiguity in the source diagrams rather than from complete extraction failure. For example, the number of patients who received treatment is sometimes only inferable from related counts, such as the number allocated to treatment and the number who did

[2] https://openrouter.ai

not receive treatment, but is not stated explicitly. Similar ambiguity appears when assigning reasons to schema fields such as other reasons or missing data. In such cases, models differ in how well they recover the intended structure. Gemini performs best overall, especially when the correct field can be inferred despite slight mismatches between the diagram wording and the target schema.

Second, all models struggle with structurally complex diagrams, including nested randomization steps and partial assessments reported for example at selected follow-up time points. These cases require both accurate parsing of the study flow and correct attachment of local counts to the appropriate fields. Performance decreases for all models on such examples, although Gemini remains the most robust.

We also observe several model-specific failure modes. Qwen3-VL 8B Instruct frequently inserts 0 into fields that are not explicitly provided instead of leaving them unspecified. While this can leave the high-level structure looking superficially plausible, it still lowers exact field accuracy and constitutes a mild hallucination, especially because the prompts explicitly discourage this behavior. In some cases, Qwen also appears to infer the correct number of study arms but fails to generate the corresponding arm objects, resulting in an empty arm list.

Finally, some errors are caused by low-level text interpretation. Subscript and superscript notation in more complex diagrams can lead to transcription mistakes. In addition, Qwen and GPT models sometimes extract full descriptive phrases instead of normalized arm or assessment names, for example `allocated to control group` instead of `control group`. These errors are relatively minor, but they reduce output consistency and make downstream normalization more difficult.

6 Future Work

CONSORT flow diagrams are complex visual summaries, and our current benchmark captures the concepts that can be most reliably standardized across studies. Future versions should expand coverage to additional structured patterns, including cluster-level versus individual-level flows, subgroup analyses versus formal re-randomization, and multi-stage designs involving multiple study-wide randomization events.

We also plan to improve evidence grounding by detecting study arms and key nodes, extracting information from localized crops, and evaluating this approach across a broader set of open-weight VLMs.

7 Conclusion

This paper presents a benchmark and capability profile study of schema-constrained CONSORT diagram extraction. Our experiments show that structured extraction from CONSORT flow diagrams is feasible with modern VLMs when grounded in a fixed schema and evaluated with structure-aware metrics. Across completed systems, the best results come from direct single-pass extraction with an expert-guided prompt, which provides the strongest balance of structural fidelity, leaf-level accuracy, and exact field recovery. The top configuration achieves a CONSORT group F1 of 0.976, a leaf micro F1 of 0.971, an exact field accuracy of 0.944, and an exact-match rate of 0.705, indicating that end-to-end extraction can recover a large fraction of complete diagrams without intermediate decomposition.

Prompting behavior depends on model capability. For the strongest proprietary models, expert-guided single-pass extraction is consistently best. For smaller open-weight models, however, single-pass settings fail on complex arm-level information, and stepwise prompting can improve key metrics by narrowing each extraction step.

These findings support decomposition, but also show limits of the current full-context stepwise design, where every pass still sees the whole diagram. A more promising next step is arm-localized decomposition: detect study arms and key nodes first, then extract from per-arm crops before merging outputs. This should reduce irrelevant context and improve local OCR and attachment decisions, especially for smaller open-weight models. At the same time, the gap between high top-line accuracy and imperfect exact-match confirms that the benchmark remains challenging and informative for capability profiling.

Acknowledgments

This research was co-funded by the European Union – European Regional Development Fund (Programme: European Funds for a Modern Economy 2021-2027, grant no. FENG.01.01-IP.02-4479/23).

References

Abdul Arbaz, Heng Fan, Junhua Ding, Meikang Qiu, and Yunhe Feng. 2024. Genflowchart: Parsing and understanding flowchart using generative ai. In *Knowledge Science, Engineering and Management: 17th International Conference, KSEM 2024, Birmingham, UK, August 16–18, 2024, Proceedings, Part I*, volume 14884 of *Lecture Notes in Computer Science*, pages 99–111. Springer.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Yufan Chen, Ching Ting Leung, Bowen Yu, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. 2025. A multi-agent system enables versatile information extraction from the chemical literature. *arXiv preprint arXiv:2507.20230*.

Google DeepMind. 2025. Gemini 3: Flash and pro models. https://deepmind.google.

Pritam Deka and Barry Devereux. 2025. Structured extraction from business process diagrams using vision-language models. *arXiv preprint arXiv:2511.22448*.

Masato Fujitake. 2024. Layoutllm: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10219–10224, Torino, Italia. ELRA and ICCL.

Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. Alpapico: Extraction of pico frames from clinical trial documents using llms. *Methods*, 226:78–88.

Google DeepMind. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-20.

Linh Hoang, Yingjun Guan, and Halil Kilicoglu. 2023. Methodological information extraction from randomized controlled trial publications: a pilot study. In *AMIA Annual Symposium Proceedings*, pages 542–551.

Sally Hopewell, An-Wen Chan, Gary S Collins, Hróbjartsson A., D. Moher, K. F. Schulz, and 1 others. 2025. Consort 2025 statement: updated guideline for reporting randomised trials. *BMJ*, 388:e081123.

Sally Hopewell, Allison Hirst, Gary S. Collins, Sue Mallett, Ly-Mee Yu, and Douglas G. Altman. 2011. Reporting of participant flow diagrams in published reports of randomized trials. *Trials*, 12:253.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2023. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. *arXiv preprint arXiv:2311.18248*.

Changkai Ji, Yang Li, Yingwen Wang, Wen He, Ying Cheng, Yuejie Zhang, Rui Feng, and Xiaobo Zhang. 2026. Evimmqa: Multimodal question answering for medical evidence extraction in systematic reviews. *Pattern Recognition*, page 113441. Available online 5 March 2026.

Daehyun Kim and Hong Yu. 2011. Figure text extraction in biomedical literature. *PLoS ONE*, 6(1):e15338.

Mathieu Laï-king and Patrick Paroubek. 2025. Evaluation of clinical trials reporting quality using large language models. *arXiv preprint arXiv:2510.04338*.

David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, PJ Devereaux, Diana Elbourne, Matthias Egger, and Douglas G Altman. 2010. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c869.

Teresa O'Leary, June Weiss, Benjamin Toll, Cynthia Brandt, and Steven L. Bernstein. 2019. Automated generation of consort diagrams using relational database software. *Applied Clinical Informatics*, 10(1):60–65.

Takamitsu Omasa, Ryo Koshihara, and Masumi Morishige. 2025. Arrow-guided vlm: Enhancing flowchart understanding via arrow direction encoding. *arXiv preprint arXiv:2505.07864*.

OpenAI. 2025. Gpt-5 technical report. https://openai.com.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. Flowlearn: Evaluating large vision-language models on flowchart understanding. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, pages 73–80. IOS Press.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118807–118833.

Pouria Rouzrokh, Bardia Khosravi, Parsa Rouzrokh, and Moein Shariatnia. 2025. Lattereview: A multi-agent framework for systematic review automation using large language models. *arXiv preprint arXiv:2501.05468*.

Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. 2024. Flowvqa: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational*

Linguistics: ACL 2024, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Simon Tannert, Marcelo G. Feighelstein, Jasmina Bogojeska, Joseph Shtok, Assaf Arbelle, Peter W. J. Staar, Anika Schumann, Jonas Kuhn, and Leonid Karlinsky. 2023. Flowchartqa: The first large-scale benchmark for reasoning over flowcharts. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 34–46, Ingolstadt, Germany. Association for Computational Linguistics.

Fan Wang, Richard L. Schilsky, David Page, Robert M. Califf, Kei Cheung, Xiaofei Wang, and Herbert Pang. 2020. Development and validation of a natural language processing tool to generate the consort reporting checklist for randomized clinical trials. *JAMA Network Open*, 3(10):e2014661.

Fei Wang, Yuewen Zheng, Qin Li, Jingyi Wu, Pengfei Li, and Luxia Zhang. 2024a. Chatschema: A pipeline of extracting structured information with large multimodal models based on schema. *arXiv preprint arXiv:2407.18716*.

Xingbo Wang, Samantha L. Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2025. Scidasynth: Interactive structured data extraction from scientific literature with large language model. *Campbell Systematic Reviews*, 21(4):e70073.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2024b. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv:2406.17755*.

Junyi Ye, Ankan Dash, Wenpeng Yin, and Guiling Wang. 2024. Beyond end-to-end vlms: Leveraging intermediate text representations for superior flowchart understanding. *arXiv preprint arXiv:2412.16420*.

A Detailed CONSORT Schema Glossary

This appendix provides a field-level description of the schema used for structured extraction from CONSORT diagrams. The goal is to make mapping decisions explicit for both annotation and evaluation.

A.1 Study-Level Structure

The study-level object captures trial-wide participant flow before arm-specific trajectories are resolved. It includes enrollment, pre-randomization filtering, randomization totals, optional post-randomization exclusions, washout structure, and global analysis-population records.

A.2 Arm-Level Structure

Each arm object captures the participant trajectory within one trial arm. The structure covers allocation, treatment receipt/completion, follow-up losses, analysis populations, exclusions, and optional internal branching such as nested randomization.

A.3 Expert Definition Alignment and Synonym Guidance

To improve annotation consistency and prompt robustness, we aligned key schema fields with expert-authored definitions and synonym inventories from our domain guidance artifact. Tables 4 and 5 summarize the mappings used during data preparation and error analysis.

Figures 3 and 4 provide structural companions to Tables 4 and 5, visualizing study-level and arm-level hierarchy components.

The expert definitions emphasize core enrollment and attrition fields, while advanced structural components in our schema (notably `partial_assessments`, `nested_randomizations`, and washout-specific branch pairing via `arms_after_washout`) require additional schema-specific normalization rules beyond direct synonym matching. We retain these components because they are required for full structural fidelity in complex CONSORT diagrams.

B Exploratory Data Analysis of Dataset

To complement the field glossary, we report corpus-level exploratory analysis of the 200-example dataset. This links schema semantics to empirical prevalence and sparsity patterns, and clarifies where extraction reliability is most sensitive to structural complexity.

B.1 Dataset Shape and Structural Prevalence

The benchmark contains $N = 200$ diagrams. The number of top-level arms per diagram is typically small (mean 2.33, median 2, range 0–9), and most examples are two-arm trials ($155/200$, 77.5%). Washout structure is uncommon (`is_washout`=true in $7/200$, 3.5%), with non-empty `arms_after_washout` in $8/200$ (4.0%). Partial assessment structures appear in $84/200$ diagrams (42.0%), while nested randomization is rare ($4/200$, 2.0%).

B.2 Field Coverage and Sparsity by Schema Level

Table 6 summarizes non-null prevalence for core schema fields.

C Additional Correctness Metrics

For detailed error analysis, we additionally report a broader set of diagnostic metrics in the appendix. Overall, the appendix metrics provide a more fine-grained view of model behavior and failure modes.

Structure-Specific Metrics. Arm inventory F1 evaluates whether the model identifies the correct set of top-level arms. Washout structure accuracy measures correct washout detection and correct matching of pre-washout and post-washout branches. Nested randomization accuracy evaluates whether internal randomization steps and their child arms are correctly represented.

Semantic Subgroup Metrics. Reason group F1 focuses on structured reason extraction, including ineligibility, non-receipt, discontinuation, and assessment-level reason groups. Analysis population F1 evaluates study-level and arm-level analysis populations, including named populations such as intention-to-treat, modified intention-to-treat, per-protocol, and other explicitly labeled analyses.

Numeric Error Metrics. For count-valued leaves, we report mean absolute error (MAE), and symmetric mean absolute percentage error (sMAPE):

$$\texttt{count_mae} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|\,, \tag{1}$$

$$\texttt{count_smape} = \frac{1}{N}\sum_{i=1}^{N} \frac{2\,|\hat{y}_i - y_i|}{|y_i| + |\hat{y}_i|}. \tag{2}$$

We summarize metrics from last three paragraphs in Table 7.

Edit-Based Diagnostics. We also summarize edit-operation statistics over the full validation set, including scalar count edits, scalar label edits, object insertions/deletions, reason insertions/deletions, and branch-repair operations. These diagnostics help characterize the types of errors made by each system and complement the aggregate quantitative metrics. We present them in Table 8.

Field	Expert definition	Synonym cues	Disambiguation rule
patients assessed for eligibility	All individuals screened/assessed for potential enrollment before allocation.	screened for eligibility; approached; assessed for enrolment/enrollment	Do not substitute enrolled or randomized counts when screening is unreported.
patients excluded before randomization total	All screened individuals not advancing to randomization for any pre-randomization reason.	patients excluded; ineligible; not enrolled; excluded during screening	Keep this as an aggregate pre-randomization total; do not merge post-randomization losses.
ineligibility.inclusionCriteriaNotMet	Excluded because inclusion criteria were not satisfied.	did not meet inclusion criteria; not eligible	Use only criterion-failure cases, not refusal or administrative reasons.
ineligibility.declinedToParticipate	Screened/approached individuals who declined consent/participation.	refused to participate; patient refusal; declined consent	Do not collapse into generic `other_reason` when refusal is explicit.
ineligibility.screeningFailure	Screening-stage failures preventing enrollment.	screening failure; not qualified at screening	Reserve for explicit screening-stage failures; not broader non-completion after randomization.
ineligibility.otherReason	Pre-randomization exclusions outside canonical categories.	other reasons; miscellaneous exclusions	Use only when no canonical ineligibility subtype applies.
patients randomized	Participants formally assigned by random allocation (including crossover sequences).	randomly assigned; underwent randomization; allocated to sequence	Prefer explicit randomization totals; avoid deriving from downstream treatment counts alone.
post randomization pre allocation exclusions	Participants excluded after randomization but before arm-specific allocation/treatment nodes.	excluded after randomization before allocation; withdrew before allocation; post-randomization exclusions	Use only for explicit between-stage losses; do not merge with pre-randomization exclusions or arm-level attrition.
is washout	Boolean indicator that a dedicated washout phase exists as a structural node.	washout period; washout phase; post-period washout	Set true only when washout is represented structurally in the flow, not when mentioned narratively.
statistical analysis	Study-level analyzed populations reported outside arm branches (e.g., ITT, mITT, safety, efficacy).	intention-to-treat; modified ITT; per-protocol; safety set; efficacy analysis set	Record only global analysis sets here; arm-specific analysis populations belong in arm objects.
arms	Ordered list of top-level arm flow objects before washout, or the sole arm list when no washout exists.	intervention arm; control arm; treatment group; sequence arm	Preserve diagram branch order to maintain deterministic alignment and downstream comparisons.
arms after washout	Ordered list of post-washout arm flows aligned one-to-one with pre-washout arm order.	post-washout arm; after washout sequence	Populate only when explicit washout splitting is present; keep positional correspondence with pre-washout arms.

Table 4: Study-level mapping between canonical schema fields and expert-authored definitions/synonym cues.

Canonical field	Condensed expert definition	High-value synonym cues	Disambiguation rule
arm name	Arm label as reported in the diagram (e.g., intervention, control, sequence name).	intervention arm; control arm; placebo arm; sequence AB/BA	Normalize superficial formatting only; preserve source arm semantics and role distinctions.
patients allocated to intervention	Participants assigned to an arm at allocation/randomization branch entry.	assigned to intervention; randomized per arm; allocated to sequence	Do not replace with received/completed treatment counts.
patients received allocated intervention	Allocated participants who initiated the assigned intervention.	received intended treatment; initiated assigned treatment	Exclude participants who never started treatment.
patients did not receive allocated intervention	Allocated participants who never initiated assigned treatment.	did not receive allocated intervention; never started treatment	Do not combine with treatment discontinuation after initiation.
patients completed treatment	Participants completing planned treatment course.	completers; completed protocol; completed intervention	Completion implies initiation; do not map follow-up completion to this field.
patients did not complete treatment	Participants who started but did not complete treatment.	drop-outs; withdrew from treatment; did not finish treatment	Keep distinct from non-receipt and analysis-stage exclusion.
discontinued intervention.*	Reason-structured treatment discontinuation after initiation.	due to adverse events; lack of efficacy; protocol violation; withdrawal	Attach only reasons linked to discontinuation nodes for that arm.
lost to follow up for primary outcome	Started treatment but missing primary-outcome follow-up data because participant could not be followed.	lost to follow-up; unable to contact; missed follow-up	Do not merge with explicit analysis exclusion categories unless diagram conflates them.
excluded from primary outcome analysis.*	Participants excluded at analysis stage with reason breakdown.	excluded from analysis; no outcome data; major protocol violation	Apply only to analysis-stage exclusions, not earlier flow attrition.
statistical analysis	Arm-specific analyzed populations (e.g., ITT, mITT, per-protocol, labeled sets).	analyzed population; ITT set; mITT set; per-protocol set; safety set	Keep distinct named populations when reported; avoid collapsing multiple analysis sets into one count.
partial assessments	Interim follow-up checkpoints (often timepoint-based) represented separately from final analysis totals.	week 12 assessed; month 6 follow-up; interim assessment	Represent each checkpoint as a separate structured object with its own count and optional reason groups.
nested randomizations	Secondary randomization branches within a top-level arm.	re-randomized; second randomization; split within arm	Create child branches only for explicit formal re-randomization steps, not ordinary subgroup reporting.

Table 5: Arm-level mapping between canonical schema fields and expert-authored definitions/synonym cues.

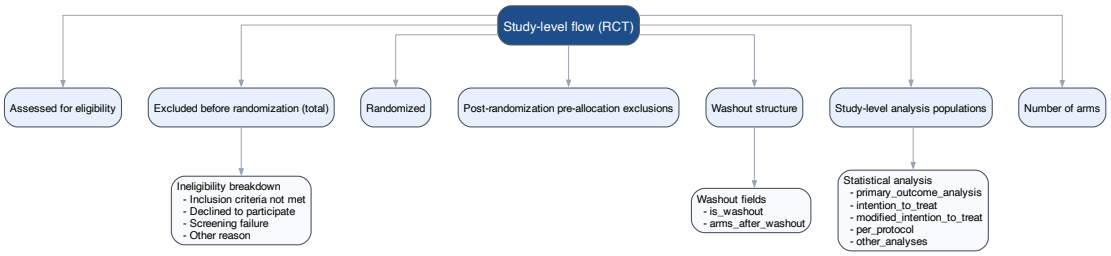


Figure 3: Study-level RCT patient-flow hierarchy aligned with Table 4.

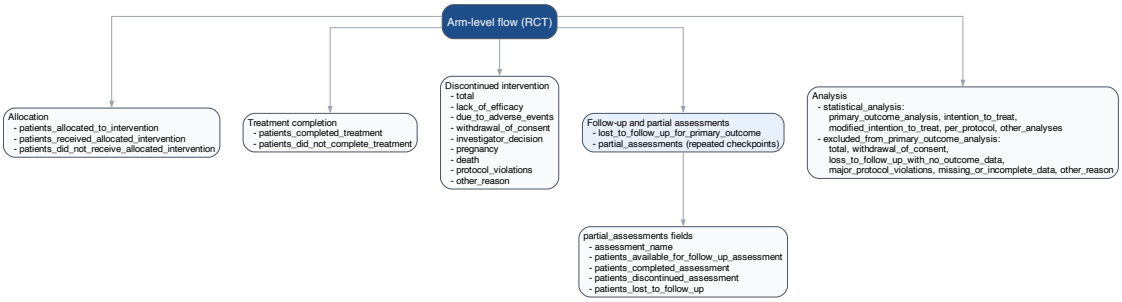


Figure 4: Arm-level RCT patient-flow hierarchy aligned with Table 5.

Study-level field	**Non-null (%)**	**Arm-level field (over all arms)**	**Non-null (%)**
patients_assessed_for_eligibility	89.0	patients_allocated_to_intervention	98.5
patients_excluded_before_randomization_total	82.5	patients_received_allocated_intervention	35.8
patients_randomized	94.5	patients_did_not_receive_allocated_intervention	23.6
post_randomization_pre_allocation_exclusions	5.0	patients_completed_treatment	14.3
statistical_analysis	9.0	patients_did_not_complete_treatment	3.0
ineligibility	83.0	discontinued_intervention	36.2
number_of_arms	100.0	lost_to_follow_up_for_primary_outcome	19.5
is_washout	100.0	statistical_analysis	63.0
		excluded_from_primary_outcome_analysis	21.2
		partial_assessments	42.0
		nested_randomizations	1.5

Table 6: Observed schema-field coverage in the 200-example gold-standard dataset used for all experiments. Study-level percentages are computed per diagram; arm-level percentages are computed over all arm objects.

Model	Setup	Study	Arm	Washout	Partial	Reasons	Analysis	Nested	MAE	sMAPE
Gemini 3 Pro	expert prompt	0.987	0.989	0.999	0.966	0.963	0.978	0.981	0.06	0.015
	basic prompt	0.975	0.933	0.998	0.797	0.863	0.904	0.963	3.10	0.042
	stepwise	0.937	0.921	0.994	0.874	0.851	0.906	0.959	8.67	0.067
Gemini 3 Flash	expert prompt	0.977	0.978	0.997	0.951	0.942	0.972	0.976	0.09	0.038
	basic prompt	0.976	0.929	0.998	0.802	0.862	0.908	0.964	3.07	0.041
	stepwise	0.953	0.925	0.993	0.874	0.856	0.906	0.959	8.64	0.037
GPT-5.1	expert prompt	0.917	0.881	0.993	0.838	0.761	0.912	0.988	2.08	0.034
	basic prompt	0.915	0.879	0.991	0.828	0.790	0.903	0.988	12.15	0.036
	stepwise	0.827	0.832	0.992	0.788	0.738	0.832	0.982	6.52	0.185
GPT-5 mini	expert prompt	0.952	0.877	0.993	0.802	0.776	0.872	0.984	9.40	0.039
	basic prompt	0.903	0.863	0.992	0.813	0.749	0.858	0.982	19.49	0.070
	stepwise	0.925	0.802	0.991	0.733	0.741	0.813	0.975	12.68	0.044
Qwen3-VL 8B Instruct	expert prompt	0.864	0.574	0.984	0.746	0.469	0.763	0.935	34.46	0.146
	stepwise	0.094	0.654	0.987	0.743	0.421	0.771	0.991	2.93	1.963
Qwen3-VL 30B A3B Instruct	expert prompt	0.807	0.407	0.967	0.671	0.428	0.499	0.991	27.16	0.280
	stepwise	0.816	0.669	0.987	0.373	0.498	0.534	0.932	19.14	0.267
Gemma-4-26B-A4B-it	expert prompt	0.624	0.654	0.987	0.743	0.363	0.640	0.991	43.27	0.602
	stepwise	0.699	0.684	0.986	0.679	0.363	0.493	0.986	48.51	0.116

Table 7: Section-level breakdown from aggregated evaluation reports on the 200-example benchmark under the same normalization and matching pipeline. Study through Nested are aggregated subsection scores, where higher is better. MAE and sMAPE are count-valued numeric error metrics, where lower is better. Partial-assessment scores are computed on the 42.0% of examples where that structure is present, and nested-randomization scores are computed on the 2.0% of examples where nested randomization appears.

Model	Setup	Count	Label	Obj+	Obj-	Reason+	Reason-	Branch	Total
Gemini 3 Pro	expert prompt	1.45	0.39	0.53	0.05	1.65	0.05	0.58	4.69
	basic prompt	4.82	1.30	1.43	0.70	5.50	0.69	2.28	16.70
	stepwise	4.28	1.77	1.77	0.38	6.12	0.37	2.33	17.01
Gemini 3 Flash	expert prompt	1.84	0.52	0.58	0.15	2.93	0.07	0.77	6.84
	basic prompt	4.78	1.35	1.47	0.72	5.62	0.69	2.34	16.96
	stepwise	4.16	1.67	1.61	0.43	5.95	0.37	2.23	16.40
GPT-5.1	expert prompt	6.26	3.18	2.28	0.79	9.00	0.42	3.57	25.48
	basic prompt	6.56	2.94	2.10	0.99	7.30	0.79	3.60	24.26
	stepwise	7.82	3.59	2.98	0.82	8.94	0.54	4.32	28.99
GPT-5 mini	expert prompt	6.75	3.30	2.43	0.68	9.87	0.50	3.39	26.90
	basic prompt	7.72	3.67	2.96	1.30	10.53	0.99	4.73	31.89
	stepwise	8.22	4.09	3.34	1.40	10.29	0.69	5.00	33.00
Qwen3-VL 8B Instruct	expert prompt	17.46	3.68	4.41	0.55	13.96	0.44	5.09	45.58
	stepwise	15.69	5.81	10.19	0.00	27.35	0.01	10.12	69.16
Qwen3-VL 30B A3B Instruct	expert prompt	21.36	3.31	3.19	1.62	12.42	0.85	5.20	47.93
	stepwise	14.53	3.68	4.08	2.66	19.05	0.86	6.84	51.68
Gemma-4-26B-A4B-it	expert prompt	13.50	5.79	10.15	0.01	27.47	0.00	10.12	67.03
	stepwise	12.25	3.14	3.38	1.08	25.73	0.02	4.53	50.12

Table 8: Mean edit operations per example from `final_metrics_jsons` on the 200-example benchmark; lower is better. Columns report scalar count edits, scalar label edits, object insertions/deletions, reason insertions/deletions, branch-repair operations, and total edit operations.

From Rules to Predictions: Federated Tabular Learning with LLM Reasoning

Afsaneh Mahanipour
Department of Computer Science
University of Kentucky
Lexington, KY, USA
ama654@uky.edu

Hana Khamfroush
Department of Computer Science
University of Kentucky
Lexington, KY, USA
khamfroush@cs.uky.edu

Abstract

Tabular data is widely used in important areas such as healthcare and finance, but building accurate models in real-world settings faces three main challenges: protecting data privacy, handling distributed data, and maintaining strong performance. Existing methods do not solve these issues together. Converting tabular data into text for Large Language Models (LLMs) can expose sensitive information, struggle with anonymized features and exact numerical values, and require expensive training while often not outperforming traditional tree-based models. In addition, many real-world datasets are spread across different institutions, making centralized training impossible. We propose a federated framework that connects distributed tabular data with LLM reasoning using decision tree rules as privacy-preserving intermediaries. Each client trains a local Random Forest and shares only extracted rules—feature comparisons and thresholds, without revealing raw data. These rules are combined into a global pool, allowing an LLM to generate a better partitioning rule without accessing any original data, adding an extra layer of privacy. Using this rule, each client learns local gradient-based corrections, which are then aggregated. We also show that this process reduces prediction error. Experiments on 12 datasets, including seven medical tasks, show that our method consistently outperforms federated baselines and achieves results close to centralized models.

1 Introduction

Tabular data, where each row is one example and each column is a feature (like numbers or categories), is widely used to make decisions in areas such as healthcare, finance, and industry (Hernandez et al., 2022; Assefa et al., 2020; Frosch et al., 2010; Borisov et al., 2022). Tasks like disease diagnosis, fraud detection, and fault prediction rely on such structured data. Due to its mixed feature types, tree-based models such as XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018) remain highly effective, often outperforming deep learning methods on tabular benchmarks (Grinsztajn et al., 2022; McElfresh et al., 2023).

The success of Large Language Models (LLMs) (Brown et al., 2020; Wei et al., 2022; Mahanipour et al., 2025; Ekanayake et al., 2025) has motivated their use in tabular prediction (Hegselmann et al., 2023; Dinh et al., 2022; Sui et al., 2024; Nam et al., 2024; Han et al., 2024; Ye et al., 2025). Most existing approaches convert table rows into text and apply LLMs via fine-tuning or prompting. However, this strategy faces three key challenges. From a data perspective, serialization can lose structural information, struggles with anonymized features, and exposes sensitive data (Sui et al., 2024; Hegselmann et al., 2023). From a model perspective, fine-tuned LLMs often fail to outperform tree-based methods despite high computational cost, while in-context learning is limited by input length (Dinh et al., 2022; Nam et al., 2024). From a distribution perspective, real-world datasets are often divided between institutions and cannot be centrally combined due to privacy regulations such as GDPR and HIPAA (Voigt and Von dem Bussche, 2017; Annas, 2003; Cohen and Mello, 2018).

Federated learning (FL) addresses this limitation by enabling multiple clients to collaboratively train models without sharing raw data (McMahan et al., 2017; Mahanipour and Khamfroush, 2025). Each client keeps data locally and shares only model updates. However, existing FL approaches for tabular data either rely on neural networks (Li et al., 2020b) or tree-based aggregation methods (Cheng et al., 2021; Li et al., 2023), and do not incorporate LLM reasoning.

The coexistence of these three tensions formulates a concrete and open research challenge: *is it possible to exploit the logical reasoning capacity of an LLM to improve tabular prediction, while*

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 970–980
July 3-4, 2026 ©2026 Association for Computational Linguistics

operating strictly within the constraints of a federated architecture, without exposing any individual training record to the model, without updating the LLM's parameters, and without sacrificing the performance advantages of tree-based ensembles on full-scale data?

In this paper, we answer this question affirmatively by proposing a federated framework that integrates LLMs into tabular prediction via logical decision tree rules as privacy-preserving intermediaries. The key insight is that a decision tree rule, a sequence of binary comparisons between feature indices and scalar thresholds, is a compact, global abstraction of the data distribution that carries no sample-level information. Rules can therefore be freely shared from clients to the server and transmitted to an LLM without exposing any individual record.

Our framework proceeds in four communication rounds. First, each client trains a local Random Forest and shares only the extracted decision rules, without exposing raw data. Second, the server aggregates these rules and uses an LLM to synthesize a refined global partitioning rule r^*, leveraging cross-client diversity while accessing no tabular data. Third, clients use r^* to model residual errors via leaf-specific predictors (a Gradient Net), sharing only model parameters. Finally, the server aggregates these parameters via FedAvg and broadcasts a global model, yielding the final prediction $F^*(\mathbf{x}) = \bar{F}(\mathbf{x}) + \eta \cdot \phi^*(\mathbf{x} \mid r^*)$.

This design addresses key challenges in privacy, modeling, and distribution. By exposing the LLM only to decision rules, which are structured and sample-free representations of feature splits, the framework avoids sharing raw feature values or labels while preserving meaningful global structure. These rules encode how the feature space is partitioned, enabling the LLM to reason about patterns and interactions without access to individual data points. The LLM is queried only once to produce r^*, after which all learning is performed locally using standard models, eliminating the need for fine-tuning or repeated inference. Throughout the process, no raw data leaves any client, and the server receives only rule text and model parameters, both of which are non-invertible to individual records. Moreover, aggregating rules across clients enriches the LLM's input with diverse, cross-client feature interactions that are not observable locally, resulting in a more informative global partition and improved predictive performance.

The main contributions of this work are as follows:

- We introduce a unified three-challenge framing for LLM-based tabular prediction, identifying data serialization, model scaling, and data silo limitations as jointly critical barriers to practical deployment in high-stakes settings where tabular data is prevalent.
- We show that decision tree rules provide a natural and effective interface between federated tabular data and LLM reasoning. They are expressive enough to capture structural feature interactions, compact enough to fit within LLM context limits, and inherently privacy-preserving since they avoid exposing raw samples.
- We propose a novel four-round federated framework in which clients transmit only rule text and leaf model parameters. The LLM refines a cross-client rule pool into a global partition r^*, and a Gradient Net trained on this rule produces sample-level error correction. Final aggregation is performed via FedAvg weighted by dataset size.
- We establish a theoretical guarantee showing that the proposed error correction step strictly decreases the expected loss for any positive step size η, providing convergence support for the framework.
- We conduct extensive experiments on eleven binary classification datasets spanning clinical, financial, and biological domains, comparing against strong federated and centralized baselines. Results show that our method achieves state-of-the-art performance among privacy-preserving approaches while remaining competitive with centralized models, demonstrating only a modest performance trade-off relative to its privacy benefits.

2 Proposed Method

We propose a novel federated framework for tabular prediction that integrates the logical interpretability of decision tree rules with the reasoning capabilities of Large Language Models (LLMs), while preserving strict data privacy across distributed clients. The key observation motivating our approach is that decision tree *rules*, sequences of axis-aligned

binary comparisons between feature indices and scalar thresholds, are compact global abstractions of a data distribution rather than representations of individual samples. Because rules convey no sample-level information, transmitting them to a central server introduces negligible privacy risk. We exploit this property to design a four-round federated protocol in which raw data never crosses any client boundary, the LLM receives no tabular rows, and all inter-client coordination is mediated through structural rule text and aggregated leaf-level model parameters.

2.1 Problem Formulation

Consider M clients, where each client i holds a private local dataset $D_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$, with $x_{ij} \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_{ij} \in \mathcal{Y}$. The feature vector x_{ij} may contain numerical, categorical, or mixed attributes; we denote its numerical and categorical components as $x^{(\text{num})}$ and $x^{(\text{cat})}$, respectively.

Each client partitions its dataset into training, validation, and test subsets, denoted by $\mathcal{D}_i^{\text{train}}$, $\mathcal{D}_i^{\text{val}}$, and $\mathcal{D}_i^{\text{test}}$. A shared global test set $\mathcal{D}^{\text{test}}$ is used for final evaluation. The total number of samples across all clients is $N = \sum_{i=1}^{M} n_i$.

The goal is to learn a global predictive model $G : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[Loss(G(x), y)\big],$$

while ensuring that no client transmits raw feature values or labels to any external party.

The framework supports binary classification ($\mathcal{Y} = \{0, 1\}$), multiclass classification ($\mathcal{Y} = \{1, \ldots, c\}$), and regression ($\mathcal{Y} = \mathbb{R}$). The loss function $Loss$ is defined as mean squared error for regression and cross-entropy for classification.

2.2 Preliminaries

Decision Trees and Rule Representation: A decision tree (Loh, 2011) recursively partitions the input space into disjoint axis-aligned regions through binary splits of the form $x_{\ell_j} \leq \tau_j$, where ℓ_j specifies the feature index and τ_j the splitting threshold at internal node j along the root-to-leaf path. For a given partitioning rule r defined by index vector $\boldsymbol{\ell}$ and threshold vector $\boldsymbol{\tau}$, the prediction function is:

$$f(x \mid \mathcal{D}^{\text{train}}, r) = \sum_{l=1}^{Node(r)} \lambda_l \cdot 1(x \in L_l(r)), \quad (1)$$

where $Node(r)$ is the number of leaf nodes, $L_l(r)$ is the l-th leaf region, $1(\cdot)$ is the indicator function, and λ_l is the leaf-specific prediction value, the empirical class distribution for classification or the mean label for regression.

The textual representation of a rule r is a sequence of human-readable if-else comparisons, e.g. `if feature_3 ≤ 0.72 then ...`. This representation contains only feature *indices* (anonymized, such as `feature_3`) and numeric *thresholds*, with no association to individual training samples.

2.3 Federated Learning

Federated learning (FL) is a distributed learning paradigm where multiple clients collaboratively train a shared global model without sharing raw data (McMahan et al., 2017; Yang et al., 2019; Li et al., 2020a). Each client i keeps a private dataset $\mathcal{D}_i$ and only shares model updates or aggregated statistics. The global objective is to minimize the weighted sum of local losses over all clients, where each client trains on its own data while contributing to a shared model. This setting enables learning in domains where data cannot be centralized due to privacy, regulatory, or institutional constraints, such as healthcare, finance, and industrial systems (Rieke et al., 2020; Yang et al., 2019; Li et al., 2020a).

A central challenge in FL is statistical heterogeneity, where client data distributions are non-IID. This can lead to client drift, where local updates diverge and slow or destabilize global convergence (Li et al., 2020b; Zhao et al., 2018). Methods such as FedProx (Li et al., 2020b), FedNova (Wang et al., 2020), and SCAFFOLD (Karimireddy et al., 2020) mitigate this issue through regularization, normalization, or variance reduction. While most FL methods focus on neural networks, recent work extends FL to tree-based models for tabular data, such as SecureBoost (Cheng et al., 2021) and FedTree (Li et al., 2023). However, these approaches rely purely on gradient statistics and do not incorporate LLM-based reasoning. Moreover, despite avoiding raw data sharing, FL systems remain vulnerable to privacy attacks on gradients and parameters (Zhu et al., 2019; Melis et al., 2019). In this work, we address these limitations by using decision tree rules as a privacy-preserving abstraction that enables LLM-guided global reasoning without exposing individual data samples.

2.4 Round 1: Local Federated Ensemble Construction

A single decision tree trained on a local $\mathcal{D}_i^{\text{train}}$ is prone to overfitting due to limited sample size (Ho, 1998). Ensemble methods mitigate this by training multiple trees on different sub-samples of the training data and averaging their predictions. In the federated setting, each client independently constructs such an ensemble, which simultaneously improves local accuracy and yields a richer pool of decision rules for subsequent cross-client synthesis.

Local ensemble. Each client i trains a Random Forest (Breiman, 2001) comprising K CART experts $\{f_k^{(i)}\}_{k=1}^K$, each trained on an independently drawn bootstrap sub-sample $\mathcal{D}_{i,k}^{\text{train}} \subset \mathcal{D}_i^{\text{train}}$ with corresponding decision rule $r_{i,k}$. The local ensemble prediction is:

$$F_i(x) \;=\; \frac{1}{K}\sum_{k=1}^{K} f_k^{(i)}\big(x \,\big|\, \mathcal{D}_{i,k}^{\text{train}},\, r_{i,k}\big)\,, \quad (2)$$

and the complete local rule set is $R_i = \{r_{i,k}\}_{k=1}^K$.

Communication. After training, client i exports the textual representation of R_i and transmits it to the server along with the scalar n_i. No raw samples, no labels, and no computed predictions are shared at this stage. The transmitted rule text is of size $O(K \cdot D_{\max})$ where $D_{\max}$ is the maximum tree depth, making communication negligible.

2.5 Round 2: Cross-Client Rule Synthesis via LLM

Global rule aggregation. The server collects all local rule sets and forms the global rule pool:

$$R^{global} \;=\; R_i[1] \cup R_i[2] \cup \cdots \cup R_i[M], \quad (3)$$

which contains $M \cdot K$ trees sourced from M statistically independent data shards. Unlike a single-client rule set, R^{global} captures feature interactions visible only within each local distribution, providing the LLM with a *cross-shard* view of the feature space that is richer and more diverse than any individual client could produce.

LLM-based rule refinement. Standard ensemble methods treat each tree in R^{global} as an independent voter and discard the structural relationships among them. We instead leverage the logical reasoning ability of an LLM to analyze and synthesize the full rule set into a single, improved partitioning rule.

The server constructs a structured prompt:

$$p \;=\; p_{\text{meta}} \oplus p_{\text{rule}} \oplus p_{\text{req}}, \quad (4)$$

where p_{meta} encodes task metadata (task type, number of features, training size), p_{rule} encodes the textual representation of R^{global}, and p_{req} specifies the required output format. Crucially, p_{rule} contains only structural information, feature indices and numeric thresholds, and no raw feature values or sample labels. The LLM is therefore never exposed to individual training records.

The server queries the LLM with p to obtain a refined global rule:

$$r^* \;=\; \text{LLM}(p) \;=\; \text{LLM}(p_{\text{meta}} \oplus p_{\text{rule}} \oplus p_{\text{req}})\,. \quad (5)$$

To reduce stochasticity in the LLM output, the server issues T independent queries (default $T = 10$), retaining the set $\{r_t^*\}_{t=1}^T$ for ensemble averaging in the final prediction stage. All T rules are broadcast to every client.

Why LLM synthesis improves upon ensemble voting. Random Forest ensembles the outputs of all rules in R^{global} but ignores the inherent structural relationships and interactions among those rules. As $M \cdot K$ grows large, analyzing these independent trees becomes increasingly difficult, and the rule set as a whole loses interpretability. The LLM addresses this by performing a form of *global rule reasoning*: it identifies features that appear consistently important across shards, abstracts away split-specific noise, and synthesizes a single globally coherent partition r^* that groups statistically similar samples into the same leaf node. We verify this empirically by measuring the average intra-leaf sample distance over all leaf nodes partitioned by r^*, which is lower than that of any single rule $r \in R^{global}$.

2.6 Round 3: Federated Local Gradient Net Training

The refined rule r^* serves a dual purpose: it provides an improved standalone partitioning of the feature space, and it guides the construction of a *Gradient Net*, a collection of leaf-specific regression models that learn to predict the residual errors of the local ensemble F_i. This design is motivated by the classical gradient boosting insight (Friedman, 2001) that predicting residual errors is often simpler than predicting ground-truth labels directly.

Local gradient sets. After receiving r^*, each client i computes the negative gradient of $Loss$ with respect to the local ensemble output for every training sample:

$$D_i^{\nabla} = \Big\{ \big(x,\ -\nabla_{F_i(x)} Loss(F_i(x), y)\big) \mid (x, y) \in D_i^{train} \Big\} \tag{6}$$

The negative gradient $-\nabla_{F_i(x)} Loss$ points in the direction of steepest loss decrease, the direction in which adjusting $F_i(x)$ would most reduce prediction error. For binary classification, $-\nabla_{F_i(x)} Loss \in \mathbb{R}$; for c-class classification, $-\nabla_{F_i(x)} Loss \in \mathbb{R}^c$.

Definition 1 (Local Gradient Set)**.** D_i^{∇} is the set of training samples paired with the negative gradients of $Loss$ evaluated at the local ensemble predictions $F_i(x)$. It differs from D_i^{∇} only in the supervision signal: labels y are replaced by the corresponding negative gradients.

Leaf-specific model fitting. The rule r^* partitions the local training data of client i into $Node(r^*)$ disjoint leaf regions $\{L_l(r^*)\}_{l=1}^{Node(r^*)}$. Within each leaf l, client i fits a local mapping function $\phi_l^{(i)}(\,\cdot\,; \theta_l^{(i)}) : \mathcal{X} \to \mathbb{R}$ (or $\mathbb{R}^c$) that approximates the negative gradient:

$$\min_{\theta_l^{(i)}} \sum_{\substack{(x,y)\in\mathcal{D}_i^{\text{train}} \\ x\in L_l(r^*)}} \Big\| \phi_l^{(i)}(x; \theta_l^{(i)}) - \big(-\nabla_{F_i(x)} Loss(F_i(x), y)\big) \Big\|_2^2 \tag{7}$$

Each $\phi_l^{(i)}$ is implemented as a CART model for classification tasks and as a pre-trained tabular foundation model (TabPFN (Hollmann et al., 2025)) for regression tasks. Leaf models are trained independently and can be parallelized across leaves, incurring minimal additional runtime.

Since all leaves of r^* are defined by the same globally broadcast rule, the leaf partitioning is identical across all clients, which is essential for the aggregation in Round 4.

Communication. After training, each client serializes the leaf model parameters $\Theta_i = \{\theta_l^{(i)}\}_{l=1}^{Node(r^*)}$ and transmits them to the server. These parameters are CART node splits (or TabPFN weights) fitted to predict gradient residuals, not to reconstruct features or labels. The communication volume is $O(T \cdot Node(r^*) \cdot d)$, which is bounded since $Node(r^*) \le 30$ as enforced by the prompt constraint.

2.7 Round 4: Federated Averaging and Global Model Assembly

FedAvg over leaf models. The server aggregates the per-client leaf model parameters using a sample-size-weighted average following the FedAvg protocol (McMahan et al., 2017). For each leaf l and each LLM query t:

$$\theta_{l,t}^* = \sum_{i=1}^{M} \frac{n_i}{N} \cdot \theta_{l,t}^{(i)}, \tag{8}$$

where n_i/N is the data fraction of client i. This weighting assigns greater influence to clients with more samples, implementing the optimal aggregation under the assumption that all clients draw from the same underlying population distribution. When a client has no samples in leaf l, it contributes a zero weight for that leaf; when only one client has samples in leaf l, its local model is retained unchanged.

Global Gradient Net. The aggregated global Gradient Net is defined as:

$$\phi^*\big(x \mid \mathcal{D}^{\nabla}, r_t^*\big) = \sum_{l=1}^{Node(r_t^*)} \phi_l^*\big(x;\, \theta_{l,t}^*\big) \cdot\ 1(x \in L_l(r_t^*)) \tag{9}$$

where $\mathcal{D}^{\nabla} = \bigcup_{i=1}^{M} D_i^{\nabla}$ denotes the conceptual union of all local gradient sets (never materialized on the server). The server broadcasts the global leaf parameters $\{\theta_{l,t}^*\}$ to all clients.

Error correction and final output. Given a test sample x, the cross-client ensemble prediction is:

$$\bar{F}(x) = \frac{1}{M} \sum_{i=1}^{M} F_i(x). \tag{10}$$

The sample-specific error correction vector is:

$$\Delta_x = \frac{\eta}{T} \sum_{t=1}^{T} \phi^*\big(x \mid \mathcal{D}^{\nabla}, r_t^*\big), \tag{11}$$

where $\eta \in \mathbb{R}^+$ is a step-size hyperparameter. The final prediction steers the ensemble output in the direction of error reduction:

$$F^*(x) = \bar{F}(x) + \Delta_x. \tag{12}$$

The framework provides strong end-to-end privacy guarantees: raw data never leaves clients, the LLM operates only on structural rule logic without access to samples, transmitted model parameters are non-invertible to individual records, and only minimal metadata such as sample counts is shared. Communication is efficient, requiring only lightweight rule text and model parameters, resulting in significantly lower overhead than gradient-based federated methods. The method also handles statistical heterogeneity by leveraging cross-client rule aggregation to learn a more generalizable partition r^*, while weighted averaging ensures alignment with the global data distribution. It scales naturally with the number of clients due to parallelizable client-side computation and minimal server workload. Finally, in the single-client setting, the framework reduces to its non-federated counterpart, introducing no additional approximation beyond data partitioning.

3 Experiments

3.1 Datasets

To evaluate our framework in realistic privacy-sensitive settings, we use 12 classification datasets from biomedical and general domains, where data are divided in a non-IID manner across clients. All datasets share a common structure: tabular data with mixed numerical and categorical features. They cover diverse medical tasks, including heart disease (HF), diabetes (ECD), cancer (LC), liver disease (LI), and hepatitis C (HE). To ensure consistency, we apply a unified preprocessing pipeline: numerical features are standardized, categorical features are ordinarily encoded with out-of-vocabulary handling, and missing values are imputed (mean for numerical, special token for categorical) (Gorishniy et al., 2021). For federated simulation, each dataset is split into $M = 3$ clients by partitioning the training data into non-IID shards.

3.2 Comparison Baseline Methods

We compare the proposed method against six representative baselines, including federated privacy-aware and centralized LLM-based approaches.

FedAvg (McMahan et al., 2017) is the standard federated learning baseline. Each client trains a two-layer MLP (hidden size 256, ReLU, dropout 0.1) for 5 local epochs per round, followed by sample-size-weighted aggregation. We run 30 rounds with early stopping. It serves as the primary benchmark for federated tabular learning. Federated-FeatLLM adapts FeatLLM (Han et al., 2024) to the federated setting. Each client queries an LLM with up to five serialized samples to generate feature transformations, which are aggregated and applied across clients. Local logistic regression models are then trained and aggregated via FedAvg. Unlike our method, it shares raw samples, making it a key comparison for evaluating rule-based versus sample-based LLM interaction. Centralized-MLP is a two-layer MLP trained on pooled data using Adam (lr 10^{-3}, weight decay 10^{-4}) for up to 200 epochs with early stopping. It represents a non-private upper bound for neural models.

Table 1: Properties of the twelve benchmark datasets used in our experiments. "#Num" and "#Cat" denote the numbers of numerical and categorical features. "IR" denotes the imbalance ratio (majority:minority class count).

ID	Dataset	#Train	#Test	#Num	#Cat	IR	#Class
HF	Heart Failure (Detrano et al., 1989)	242	61	13	0	1.17	2
LC	Lung Cancer (Raut et al., 2021)	228	57	0	15	2.27	2
ECD	Early Diabetes (Islam et al., 2019)	416	104	2	14	1.59	2
LI	Indian Liver (Ramana et al., 2012)	464	117	9	1	2.50	2
HE	Hepatitis C (Hoffmann et al., 2018)	474	119	9	1	9.07	2
PID	Pima Diabetes (Smith et al., 1988)	614	154	8	0	1.87	2
FH	Framingham (O'Donnell and Elosua, 2008)	3390	848	8	5	5.44	2
BL	Blood (Yeh, 2008)	598	150	4	0	3.20	2
CR	Credit (Hofmann, 1994)	800	200	7	13	2.33	2
BA	Bank (Moro et al., 2014)	28934	9043	7	9	7.55	2
AD	Adult (Kohavi, 1996)	39074	9769	6	8	3.17	2
JA	Jannis (Gorishniy et al., 2021)	66987	16747	54	0	2.08	4

For centralized LLM-based methods, which require full access to pooled data and are evaluated on five benchmarks using accuracy, we consider three approaches. LIFT (GPT-3.5) (Dinh et al., 2022) fine-tunes GPT-3.5 on serialized tabular data, requiring both parameter updates and raw data access. TP-BERTa (RoBERTa) (Sui et al., 2024) pre-trains a tabular-specific language model and fine-tunes it on each dataset, achieving strong performance but requiring full data access. FeatLLM (GPT-4o) (Han et al., 2024) generates feature transformations from serialized samples without fine-tuning, but still exposes raw data, unlike our rule-based approach. Our method also uses GPT-4o, but only over rule-based representations, avoiding any exposure of raw samples.

3.3 Evaluation Metrics

In this work, we report multiple complementary metrics. **Accuracy** provides overall correctness for comparability. **AUC** evaluates ranking quality in a threshold-independent manner, which is particularly important under imbalance. **F1-score**, the

Table 2: Performance comparison across multiple datasets using six evaluation metrics (Accuracy, Precision, Recall, F1-score, MCC, and AUC).

Dataset	Method	Acc	Precision	Recall	F1-score	MCC	AUC
Heart-Failure	FedAvg	0.8478	**0.8767**	0.8421	0.8591	0.6945	0.9157
	Federated-FeatLLM	0.8261	0.8421	0.8421	0.8421	0.6486	0.8882
	Centralized-MLP	0.8551	0.8684	0.8684	0.8684	0.7071	0.9028
	Ours	**0.8695**	0.8718	**0.8947**	**0.8831**	**0.7359**	**0.9183**
Lung-Cancer	FedAvg	0.9149	0.9512	0.9512	0.9512	0.6179	0.9431
	Federated-FeatLLM	0.9149	0.9523	0.9512	0.9514	0.6081	0.9634
	Centralized-MLP	0.8723	0.907	0.9512	0.9286	0.3403	0.9593
	Ours	**0.9362**	**0.975**	0.9512	**0.963**	**0.7354**	**0.9756**
Early-Diabetes	FedAvg	0.9494	0.9592	0.9592	0.9592	0.8925	0.9946
	Federated-FeatLLM	0.9367	0.94	0.9592	0.9495	0.8651	0.9816
	Centralized-MLP	0.962	0.96	0.9796	0.9697	0.9192	0.9939
	Ours	**0.9747**	**0.9796**	**0.9796**	**0.9796**	**0.9463**	**0.9973**
Liver	FedAvg	0.7386	0.7778	0.8889	0.8296	0.291	0.7771
	Federated-FeatLLM	0.7273	0.7746	0.873	0.8209	0.2662	**0.7879**
	Centralized-MLP	0.7386	**0.7857**	0.873	0.8271	**0.3052**	0.7803
	Ours	**0.7386**	0.75	**0.9524**	**0.8392**	0.239	0.76
Pima	FedAvg	0.75	0.7143	0.4878	0.5797	0.4258	**0.8426**
	Federated-FeatLLM	0.7241	0.6552	0.4634	0.5429	0.3644	0.8007
	Centralized-MLP	0.7328	0.6786	0.4634	0.5507	0.3836	0.8338
	Ours	**0.7672**	**0.75**	**0.5122**	**0.6087**	**0.4679**	0.828
Framingham	FedAvg	0.8399	0.3529	0.0619	0.1053	0.0925	0.6867
	Federated-FeatLLM	0.8399	0.3529	0.0619	0.1053	0.0925	**0.7062**
	Centralized-MLP	0.8477	0.5	**0.0619**	**0.1101**	**0.1341**	0.6905
	Ours	**0.8493**	**0.5**	0.0103	0.0202	0.0543	0.7032
Hepatitis-C	FedAvg	0.957	0.9091	0.7692	0.8333	0.8126	0.9962
	Federated-FeatLLM	0.9462	1.0	0.6154	0.7619	0.761	0.9962
	Centralized-MLP	0.9677	1.0	0.7692	0.8696	0.8611	1.0
	Ours	**0.9785**	**1.0**	**0.8462**	**0.9167**	**0.9086**	**0.999**

Table 3: Accuracy comparison of centralized LLM-based methods and the proposed approach across five benchmark tabular datasets.

Method	Blood	Credit	Bank	Adult	Jannis
LIFT (GPT-3.5)	0.689	0.691	0.825	0.810	0.647
TP-BERTa (RoBERTa)	0.761	0.730	**0.916**	0.844	0.659
FeatLLM (GPT-4o)	**0.768**	0.701	0.887	0.842	0.540
Ours	0.761	**0.762**	0.898	**0.859**	**0.6653**

harmonic mean of Precision and Recall, offers a balanced view of minority-class performance, especially in highly skewed datasets.

To further characterize performance, we report **Precision** (fraction of correct positive predictions) and **Recall** (fraction of detected positives), reflecting clinical trade-offs between false alarms and missed cases. **MCC** (Chicco and Jurman, 2020) provides a robust single-score summary using all confusion matrix entries. All results are computed on a fixed test split with three random seeds, reporting mean and standard deviation.

3.4 Results and Analysis

Table 2 presents a comprehensive performance comparison of the proposed federated framework against three baseline methods across seven clinical benchmark datasets, evaluated on six complementary metrics: Accuracy, Precision, Recall, F1-Score, Matthews Correlation Coefficient (MCC), and Area Under the ROC Curve (AUC). The baselines include two federated competitors,

FedAvg (McMahan et al., 2017) and Federated-FeatLLM (Han et al., 2024), and one centralized reference, Centralized-MLP, which is trained on the fully pooled training set and therefore represents an optimistic upper bound for non-federated learning. Crucially, the proposed method operates under identical privacy constraints to FedAvg and Federated-FeatLLM (no raw data leaves any client), yet achieves performance that frequently exceeds even the centralized MLP upper bound, demonstrating that LLM-guided rule synthesis compensates for the accuracy cost of data partitioning.

The proposed method achieves the best Accuracy, F1, and MCC on 6/7 datasets, with clear gains. For example, on Early Diabetes it reaches 0.9747 Accuracy (+1.27% over Centralized-MLP, +2.53% over FedAvg), and on Hepatitis C it achieves 0.9785 Accuracy and 0.9086 MCC, outperforming both baselines. These results show that LLM-guided rule synthesis yields better feature partitions than gradient-based or local feature methods.

On Heart Failure, it leads all metrics (e.g., 0.8695 Accuracy, 0.8947 Recall), even surpassing the centralized model, with improved recall being clinically important. On Lung Cancer, it achieves top Accuracy (0.9362) and AUC (0.9756), demonstrating robustness under small, sparse data by leveraging cross-client rules.

On Early Diabetes, it attains the best results across all metrics, consistently outperforming both centralized and federated baselines. On Indian Liver, it matches best Accuracy but achieves higher Recall and F1, indicating a sensitivity–specificity trade-off in a difficult, imbalanced setting.

On Pima Diabetes, it leads most metrics, especially MCC (0.4679), reflecting better class balance handling. Finally, on Hepatitis C, it shows the strongest improvement, with higher Recall and MCC despite perfect Precision across methods, highlighting better detection of rare positives.

Table 3 compares the proposed federated method with centralized LLM-based models (LIFT, TP-BERTa, FeatLLM) on five datasets. Unlike these methods, which require full data access and expose raw samples to LLMs, the proposed approach operates under strict federated privacy without sharing data.

Despite this, it achieves competitive or better Accuracy on 4/5 datasets: it outperforms all baselines on Credit (0.762), Adult (0.859), and Jannis (0.6653), and is close on Bank (0.898 vs 0.916 TP-BERTa). On Blood, it matches TP-BERTa (0.761) with only a negligible gap to the best result. Overall, the average accuracy loss versus the best centralized model is under 1%, indicating minimal cost for strong privacy guarantees.

Compared to LIFT and FeatLLM, which directly use raw data with LLMs, the proposed method consistently performs better (beating LIFT on all datasets and FeatLLM on 4/5). This shows that using decision-tree rules as privacy-preserving intermediaries can achieve both strong privacy and competitive accuracy.

3.5 Conclusion

This paper studied how to integrate LLM reasoning into tabular prediction under federated learning constraints, where raw data cannot be shared across institutions. Existing LLM-based approaches typically rely on centralized data access or direct serialization of individual records, which is not feasible in privacy-sensitive settings. To address this, we proposed using decision tree rules as a privacy-preserving interface between distributed tabular data and LLM reasoning. Experiments on 12 standard benchmarks show consistent improvements over strong federated baselines such as FedAvg and Federated-FeatLLM. The method achieves particularly strong results on challenging medical datasets, often improving key metrics like F1-score and MCC. In several cases, it also matches or outperforms centralized neural models, showing that federated learning does not necessarily require a large performance trade-off. On standard benchmarks, the proposed method remains competitive with centralized LLM-based approaches. Overall, the results demonstrate that combining federated learning with LLM reasoning through decision tree rules is an effective and practical approach for privacy-preserving tabular prediction. It enables strong performance without sharing raw data and significantly narrows the gap between federated and centralized learning.

Acknowledgement

This work is funded by career grant provided by the National Science Foundation (NSF) under the grant number 2340075.

References

George J Annas. 2003. Hipaa regulations—a new era of medical-record privacy?

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the first ACM international conference on AI in finance*, pages 1–8.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, 35(6):7499–7519.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. 2021. Secureboost: A lossless federated learning framework. *IEEE intelligent systems*, 36(6):87–98.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

I Glenn Cohen and Michelle M Mello. 2018. Hipaa and protecting health information in the 21st century. *Jama*, 320(3):6–7.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. 2022. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784.

Vinu Ekanayake, Md Sultan Al Nahian, and Ramakanth Kavuluru. 2025. Mining social media for barriers to opioid recovery with llms. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 83–99.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Dominick L Frosch, David Grande, Derjung M Tarn, and Richard L Kravitz. 2010. A decade of controversy: balancing policy with evidence in the regulation of prescription drug advertising. *American Journal of Public Health*, 100(1):24–32.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. 2024. Large language models can automatically engineer features for few-shot tabular learning. *arXiv preprint arXiv:2404.09491*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pages 5549–5581. PMLR.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45.

Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844.

Georg Hoffmann, Andreas Bietenbeck, Ralf Lichtinghagen, and Frank Klawonn. 2018. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *Journal of Laboratory and Precision Medicine*, 3(6).

Hans Hofmann. 1994. Statlog (german credit data). UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

MM Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. 2019. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019*, pages 113–125. Springer.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Ron Kohavi. 1996. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 96, pages 202–207.

Qinbin Li, Wu Zhaomin, Yanzheng Cai, Ching Man Yung, Tianyuan Fu, Bingsheng He, and 1 others. 2023. Fedtree: A federated learning system for trees. *Proceedings of Machine Learning and Systems*, 5:89–103.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Wei-Yin Loh. 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23.

Afsaneh Mahanipour, Abdullah-Al-Zubaer Imran, and Hana Khamfroush. 2025. Federated reprogramming knowledge distillation for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 143–152. Springer.

Afsaneh Mahanipour and Hana Khamfroush. 2025. Embedded federated feature selection with dynamic sparse training: balancing accuracy-cost tradeoffs. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. 2023. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36:76336–76369.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. Pmlr.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE.

Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. 2024. Optimized feature generation for tabular data via llms with decision tree reasoning. *Advances in neural information processing systems*, 37:92352–92380.

Christopher J O'Donnell and Roberto Elosua. 2008. Cardiovascular risk factors. insights from framingham heart study. *Revista Española de Cardiología (English Edition)*, 61(3):299–310.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Bendi Venkata Ramana, M Surendra Prasad Babu, and NB Venkateswarlu. 2012. A critical comparative study of liver patients from usa and india: an exploratory analysis. *International Journal of Computer Science Issues (IJCSI)*, 9(3):506.

Smita Raut, Shraddha Patil, and Gopichand Shelke. 2021. Lung cancer detection using machine learning approach. *International Journal of Advance Scientific Research and Engineering Trends (IJASRET)*.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, and 1 others. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119.

Jack W Smith, James E Everhart, William C Dickson, William C Knowler, and Robert Scott Johannes. 1988. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Hangting Ye, Jinmeng Li, He Zhao, Dandan Guo, and Yi Chang. 2025. Llm meeting decision trees on tabular data. *arXiv preprint arXiv:2505.17918*.

I-Cheng Yeh. 2008. Blood transfusion service center. UCI Machine Learning Repository. https://doi.org/10.24432/C5GS39.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.

MedBench: Deliberative Evaluation of Medical Language Models

Pratik Jalan[*,1] **Mukul Joshi**[*,1] **Akhilesh Magotra**[2] **Kshitij Jadhav**[1]
[1]Indian Institute of Technology Bombay, Mumbai, India
[2]Birla Institute of Technology & Science, Pilani, India
{30006531,30004808,kshitij.jadhav}@iitb.ac.in akhilesh.magotra@gmail.com

Abstract

We introduce **MedBench**, a benchmark for evaluating medical language models as deliberating agents rather than isolated predictors. MedBench evaluates eight models (4B-32B) on 19,625 questions from six medical QA datasets using Consensus-Aware Model Panel (CAMP), a two-tier protocol in which five 4B–8B models answer independently, revise after observing peer reasoning, and escalate persistent disagreements to larger 20B-32B models. Compared with zero-shot, few-shot, and chain-of-thought baselines, CAMP shows that deliberation is not uniformly accuracy-improving, but reveals interaction-driven behaviors hidden by single-model evaluation. On PubMedQA without external context, the 4B-8B panel outperforms the evaluated 20B-32B individual zero-shot models (54.1% vs. 33.9%), and achieves the best evaluated result with context (75.7%), suggesting that structured interaction can sometimes complement scale. Across five datasets, initial inter-model agreement is positively associated with correctness and serves as a useful difficulty signal. However, on MedXpertQA, unanimous agreement yields only 6.6% accuracy despite 14.4% overall accuracy, suggesting correlated ignorance, where shared biases make consensus misleading. Error analysis shows that most failures are debate-insufficient cases, where incorrect majorities persist despite interaction (93-97%), while debate-harmful cases account for 3-7%. MedBench positions deliberative evaluation as a complement to accuracy-centric benchmarking, measuring when model interaction corrects errors, reinforces shared mistakes, or signals the need for stronger evidence and human review.

1 Introduction

The dominant paradigm in medical AI benchmarking is simple: ask a model a question, record whether it answers correctly, and report a single accuracy number. Considerable progress has followed from scaling (Nori et al., 2023; Sellergren et al., 2026; Singhal et al., 2025) and prompting (Wei et al., 2022; Brown et al., 2020). Yet this paradigm has a structural blind spot: it cannot detect failure modes that only emerge when models are placed in relation to each other, including shared biases, correlated misconceptions, and the qualitative difference between "every model is wrong because the question is hard" and "every model agrees because they all share the same hallucination."

Medicine itself provides a precedent for deliberative evaluation. Physicians facing genuine diagnostic uncertainty consult colleagues, revise initial impressions in light of contradicting reasoning, and escalate to specialists when the collective remains uncertain. Benchmarks, by contrast, ask language models the same question in isolation. The result is that two qualitatively different failure modes are collapsed into a single metric: cases where models are wrong but could be corrected through interaction, and cases where models are wrong in a way that collective deliberation would only reinforce the incorrect reasoning. We operationalize this intuition in **MedBench**, a benchmark that evaluates medical language models not as isolated oracles but as deliberating agents. MedBench specifies both a collection of evaluation datasets spanning the clinical difficulty gradient and the CAMP protocol as its assessment mechanism; the two are designed together so that the protocol's deliberative structure is what makes the benchmark's diagnostic findings possible.

The CAMP (Consensus-Aware Model Panel) is an evaluation protocol, not a deployment proposal. In this protocol, models iteratively revise their answers after observing peer reasoning under a fixed update protocol (Du et al., 2023; Liang et al., 2023). Unlike self-consistency (Wang et al., 2023) and ensemble voting, which aggregate independent pre-

[*]Equal contribution.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 981–991
July 3-4, 2026 ©2026 Association for Computational Linguistics

dictions, CAMP enables iterative interaction between models, making the influence structure observable and measurable. For each question, the protocol generates an inter-model agreement trajectory, per-model position-change sequences, and a resolution pathway, collectively constituting a richer characterization of model capability than accuracy alone. We treat deliberation as an observable process that reveals latent error structure in language models: interaction converts hidden error correlations into measurable signals.

Why this matters in medicine: In high-stakes domains such as clinical decision support, understanding when model confidence is reliable versus when it reflects shared bias is not merely an academic concern. A finding of unanimous model agreement on a specialist question is equally consistent with collective expertise and with collective hallucination; standard benchmarks cannot distinguish the two. MedBench provides formal tools to make this distinction, including the agreement ratio as an uncertainty proxy and the correlated ignorance diagnosis for cases where consensus signals failure rather than reliability.

The evaluation gap: Medical QA benchmarks span a wide difficulty gradient, from MMLU-style questions tractable for 4B models to ten-option specialist questions that remain difficult for the evaluated model sizes (Jin et al., 2020; Zuo et al., 2025; Hendrycks et al., 2021). Standard evaluation conflates confident correct reasoning with lucky guessing, and cannot distinguish "all models disagree because the question is hard" from "all models agree because they share the same bias." MedBench is designed to make this distinction explicit.

Contributions:

1. **MedBench benchmark.** To our knowledge, MedBench is among the first benchmarks to evaluate medical language models under a deliberative protocol across six datasets and eight models (19,625 questions), extending medical NLP evaluation beyond isolated single-model accuracy.

2. **Deliberation can complement scale in a scoped setting.** A panel of 4–8B models outperforms the individual 20–32B models evaluated here on PubMedQA without context and obtains the strongest result among our evaluated settings on PubMedQA with context (75.68%), suggesting that deliberation can sometimes complement scale.

3. **Agreement ratio as an uncertainty proxy.** Across five datasets, the fraction of Panel 1 models initially agreeing is a strong predictor of final accuracy, serving as a practical uncertainty signal. On MedXpertQA the relationship inverts completely, providing a diagnostic criterion for correlated ignorance, a failure mode where unanimous collective confidence actively misleads.

4. **Debate error taxonomy.** Debate failures partition into two primary debate-related modes: debate-insufficient (93–97%, incorrect majority persists despite interaction) and debate-harmful (3–7%, a correct minority model is overturned by peer influence), with Panel 2 insufficiency tracked separately for escalated cases. These error types are invisible to accuracy-only evaluation and suggest different intervention strategies.

2 Related Work

Medical LLM benchmarking: Foundational medical QA benchmarks include MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), and MMLU (Hendrycks et al., 2021). Performance has risen sharply with scale (Nori et al., 2023) and domain fine-tuning (Sellergren et al., 2026). MedMCQA (Pal et al., 2022) expanded evaluation to Indian medical licensing questions. MedXpertQA (Zuo et al., 2025) recently introduced ten-option specialist questions designed to resist saturation; our analysis suggests that it reveals a qualitatively different failure mode. Most benchmark evaluations still treat models in isolation; MedBench instead uses a deliberative protocol to study interaction dynamics across this landscape.

Multi-agent deliberation: (Du et al., 2023) showed that having multiple LLM instances debate their responses over multiple rounds improves factual validity and mathematical reasoning. (Liang et al., 2023) introduced structured multi-agent debate to encourage divergent thinking and found that forcing models to take opposing positions reduces groupthink. (Chan et al., 2023) applied multi-agent debate for chatbot evaluation. CAMEL (Li et al., 2023) explored role-playing agents for collaborative task completion. In clinical reasoning, (Misaghi et al., 2026) study deliberative multi-agent

councils for ophthalmology vignettes. Our work differs from these lines in three ways: (1) we use deliberation as an evaluation protocol applied to a multi-dataset medical QA benchmark suite rather than as a generation strategy for a single domain; (2) we introduce a two-tier hierarchical structure with explicit escalation; and (3) we measure diagnostic failure modes, including agreement-ratio inversion, debate-induced corruption, and escalation pathways, alongside performance gains.

Self-consistency and ensemble methods: Self-consistency (Wang et al., 2023) samples diverse reasoning paths and takes the majority vote without explicit argument exchange. MedBench's debate rounds enable explicit evidence sharing, allowing us to track social influence events (how often one model switches another's answer) and to separate productive self-correction from harmful corruption. This separation is central to our analysis of when deliberation helps and when it fails.

Confidence and calibration: (Xiong et al., 2024) systematically evaluated confidence elicitation in LLMs and found that expressed confidence is poorly calibrated under distribution shift. Our error analysis corroborates this at scale: models report 0.706 to 0.815 average confidence on incorrect answers, approaching their confidence on correct ones, and the observed high confidence on incorrect answers prevents confidence from functioning as a reliable escalation signal in our setting.

Disagreement as signal: (Kuai et al., 2026) showed that language models trained on overlapping corpora exhibit systematic behavioral entanglement that undermines the independence assumptions required for reliable ensemble voting. MedBench reframes inter-model disagreement not as noise to suppress but as an informative signal: high disagreement reliably indicates question difficulty, while unexpectedly unanimous agreement exposes correlated failure modes. This shift from disagreement-as-error to disagreement-as-diagnostic is central to MedBench's contribution as an evaluation framework.

3 Benchmark Datasets

MedBench evaluates models on six medical QA datasets designed to span a difficulty gradient from standard medical curricula to expert specialist knowledge. Table 1 provides an overview.

Dataset	N	Options	Domain
MedQA	12,723	4	USMLE clinical
MedXpertQA	2,450	10	Expert specialist
MetaMedQA	1,373	4	Multilingual clinical
MMLU-Med	1,089	4	6 med. subjects
PubMedQA	995	3	Biomed. literature
PubMedQA+C	995	3	Biomed. + abstract
Total	**19,625**		

Table 1: Composition of the MedBench benchmark suite, including dataset size, number of answer options, and clinical or biomedical domain coverage.

MedQA (Jin et al., 2020) contains four-option questions in USMLE Step 1/2/3 style, testing clinical reasoning required for US medical licensure. **MedXpertQA** (Zuo et al., 2025) presents ten-option expert-level questions spanning treatment, diagnosis, and basic science. With a random baseline of 10%, this dataset is especially challenging for the evaluated models. **MetaMedQA** aggregates multilingual clinical MCQs testing robustness across question styles and language registers. **MMLU-Med** (Hendrycks et al., 2021) draws six medical subcategories from the Massive Multitask Language Understanding benchmark. **PubMedQA** and **PubMedQA+C** (Jin et al., 2019) ask three-class (yes/no/maybe) questions about biomedical research; the paired design isolates retrieval from comprehension by providing or withholding the source abstract.

Together, these datasets span a controlled difficulty gradient, enabling us to study how deliberation behaves under regimes where the evaluated models have comparatively stronger performance (MMLU-Med, MedQA), partial evidence or mixed robustness (MetaMedQA, PubMedQA+C), and expert-level difficulty or absence of provided context (MedXpertQA, PubMedQA without context).

4 The MedBench Evaluation Protocol

The CAMP routes each question through a hierarchical two-tier deliberation protocol. Figure 1 illustrates the overall architecture.

4.1 Formal Protocol

We formalize deliberation as an iterative update process where models revise their predictions based on peer responses. This allows us to quantify not only outcomes but also interaction dynamics such as influence and leadership. The protocol can be inter-

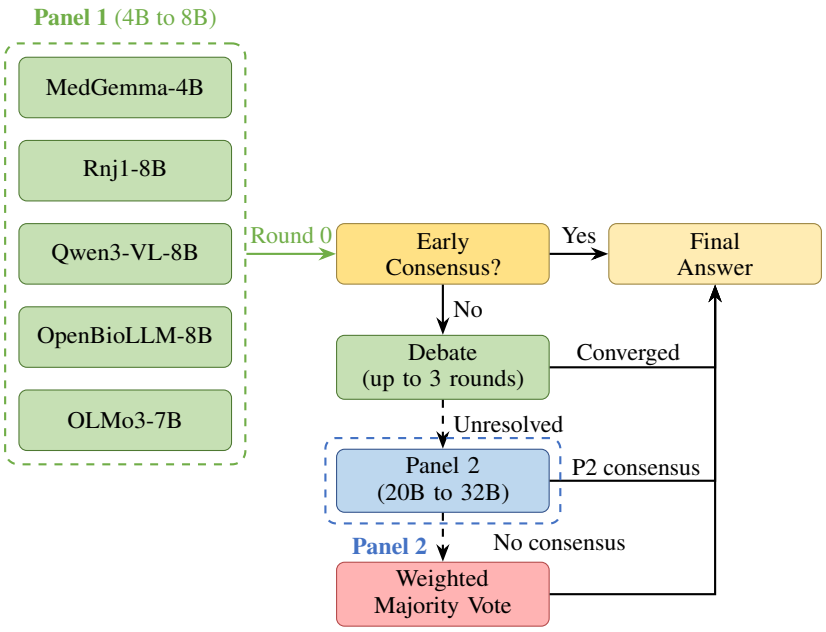


Figure 1: CAMP evaluation architecture. Panel 1 models first reason independently (Round 0). Questions where at least 4 of 5 agree are resolved immediately via Early Consensus. Those that fail are debated for up to three rounds. Persistent Panel 1 disagreement triggers escalation to Panel 2 specialist models (20B to 32B); a confidence-weighted majority vote (WMV) fallback fires only if Panel 2 also fails to converge.

preted as a discrete-time opinion dynamics system over model predictions (Cisneros-Velarde, 2024), where each model's state is updated by exposure to peer reasoning rather than by direct parameter modification.

Let $\mathcal{Q}$ be a question with answer set $\mathcal{A}$, and let $\mathcal{M}_1 = \{m_1, \ldots, m_5\}$ denote the five Panel 1 models.

Independent reasoning (Round 0): Each model m_i receives a role-structured prompt ("You are a board-certified physician on a diagnostic panel...") requiring a JSON response containing a selected option $a_i^{(0)} \in \mathcal{A}$, a scalar confidence $c_i^{(0)} \in [0, 1]$, and a free-text clinical reasoning chain $r_i^{(0)}$.

Early consensus: After Round 0, the agreement ratio (AR) is computed as:

$$\mathrm{AR}^{(t)} = \frac{\max_{a \in \mathcal{A}} \left| \{ i : a_i^{(t)} = a \} \right|}{|\mathcal{M}_1|} \tag{1}$$

If $\mathrm{AR}^{(0)} \geq 0.8$ (at least 4 of 5 models agree), the majority answer becomes the Panel 1 verdict without debate.

Structured debate: When $\mathrm{AR}^{(0)} < 0.8$, each model m_i receives the full set of peer responses from the previous round and updates its answer:

$$\left(a_i^{(t)},\, c_i^{(t)},\, r_i^{(t)}\right) = m_i\Big(\mathcal{Q},\ \big\{(a_j^{(t-1)}, r_j^{(t-1)})\big\}_{j \neq i}\Big) \tag{2}$$

Debate terminates when $\mathrm{AR}^{(t)} \geq 0.8$ or $t = R = 3$.

Influence and leadership: We define two descriptive interaction metrics for this study, using opinion-dynamics work as conceptual background rather than adopting a standard clinical metric. The influence of model m_i on m_j across all debate rounds is:

$$\mathrm{Inf}_{i \to j} = \sum_{t=1}^{R} \mathbf{1}\left[a_j^{(t)} \neq a_j^{(t-1)} \ \wedge \ a_j^{(t)} = a_i^{(t-1)} \right] \tag{3}$$

The leader-follower index for model m_i is:

$$\mathrm{LF}_i = \frac{\sum_{j \neq i} \mathrm{Inf}_{i \to j} - \sum_{j \neq i} \mathrm{Inf}_{j \to i}}{\sum_{j \neq i} \left(\mathrm{Inf}_{i \to j} + \mathrm{Inf}_{j \to i} \right) + \epsilon} \tag{4}$$

with $\epsilon = 1$ to prevent division by zero. $\mathrm{LF}_i > 0$ indicates a net debate leader; $\mathrm{LF}_i < 0$ indicates a net follower.

Panel 2 escalation: Questions where Panel 1 fails to reach consensus after $R = 3$ rounds are forwarded to Panel 2 ($\mathcal{M}_2 = \{$GPT-OSS-20B, MedGemma-27B, Qwen3-VL-32B$\}$). Each Panel

Model	Size	Panel	Type
MedGemma	4B	P1	Medical
Rnj1	8B	P1	General
Qwen3-VL	8B	P1	Multimodal
OpenBioLLM	8B	P1	Biomedical
OLMo3	7B	P1	General
GPT-OSS	20B	P2	General
MedGemma	27B	P2	Medical
Qwen3-VL	32B	P2	Multimodal

Table 2: Model configuration for the two-panel CAMP evaluation setup. Panel 1 contains the primary models and handles 94.5–99.5% of questions, while Panel 2 is invoked only under persistent Panel 1 disagreement.

Model	Panel	Source
MedGemma-4B	P1	HF
Rnj1-8B	P1	HF
Qwen3-VL-8B	P1	HF
OpenBioLLM-8B	P1	HF
OLMo3-7B	P1	HF
GPT-OSS-20B	P2	HF
MedGemma-27B	P2	HF
Qwen3-VL-32B	P2	HF

Table 3: Public Hugging Face sources for the eight models evaluated in MedBench.

2 model first reasons independently (blind to its peers' Panel 2 responses), and then participates in up to two additional debate rounds if Panel 2 initial agreement falls below 2/3. When Panel 2 also fails, confidence-weighted majority vote across all eight models provides the final answer.

Design rationale: The five-model Panel 1 is intended to balance model diversity, local deployability, and a clear majority threshold. The 4-of-5 early-consensus rule resolves questions only when the small-model panel shows strong agreement, while 3–2 splits remain eligible for debate. The three-round limit allows models to revise after seeing peer reasoning while bounding cost and avoiding open-ended deliberation; we do not claim this is the optimal number of rounds. Panel 2 is reserved for persistent Panel 1 disagreement, and weighted majority voting is used only as a final fallback when both tiers fail to converge.

4.2 Model Configuration

Table 2 lists all eight models. Panel 1 comprises five open-weight models spanning 4B to 8B parameters, each deployable locally. Panel 2 comprises three larger models spanning 20B to 32B. This separation between tiers helps us compare deliberation with parameter scale while keeping larger-model conclusions cautious because Panel 2 is invoked only for a small escalated subset. The public source table lists model identities for reproducibility (Sellergren et al., 2026; Vaswani et al., 2025; Bai et al., 2025; OpenBioLLM, 2024; Olmo et al., 2026; OpenAI et al., 2025).

5 Experimental Setup

Baselines: For each of the eight models we run: (1) zero-shot with a minimal prompt; (2) five-shot with domain-matched examples; and (3) chain-of-thought (CoT) prompting (Wei et al., 2022; Brown et al., 2020). In the CoT setting, models are prompted to reason step by step before committing to a final answer. All three protocols form the individual model baseline suite.

Metrics: Primary metric is accuracy. Secondary metrics include: debate uplift (CAMP's Round 0 majority accuracy), escalation rate, self-correction rate (fraction of position changes toward the correct answer during debate), and corruption rate (fraction away from the correct answer). We apply McNemar's two-tailed test for pairwise comparisons, with $p < 0.001$ considered significant given dataset sample sizes.

Implementation: All Panel 1 models run locally on A100 GPUs. Panel 2 models are accessed via API endpoints. All models are evaluated under identical prompting and output constraints to ensure comparability across protocols. All models use a consistent JSON response format; temperature is set to 0 for all Round 0 responses and to 0.7 for debate rounds to allow opinion updating.

6 Individual Model Results

Table 4 reports accuracy for all eight models under all three baseline protocols.

Table 4 also shows dataset-specific empirical patterns among Panel 1 zero-shot models. Qwen3-VL-8B has the highest Panel 1 zero-shot accuracy on four of six datasets, especially the four exam-style MCQ datasets: MedQA, MedXpertQA, MetaMedQA, and MMLU-Med. OpenBioLLM-8B has the highest Panel 1 zero-shot accuracy on PubMedQA without context but is weak on several MCQ exam datasets, suggesting that biomedical pretraining alone does not ensure robust option-selection behavior in this evaluation. MedGemma-

Protocol	Model	Accuracy (%)					
		MedQA	MXQA	MMed	MMLU	PMed	PMed+C
Zero-Shot	MedGemma-4B	46.01	7.77	39.08	60.17	36.15	**64.96**
	Rnj1-8B	35.26	7.54	28.86	54.72	14.74	57.37
	Qwen3-VL-8B	**53.53**	**7.89**	**47.66**	**76.10**	20.83	60.62
	OpenBioLLM-8B	7.00	3.39	19.26	42.09	**41.11**	22.21
	OLMo3-7B	31.88	6.51	27.75	58.24	12.51	49.09
	GPT-OSS-20B	60.79	19.59	56.90	75.18	33.88	68.27
	MedGemma-27B	57.92	11.72	55.35	77.13	28.99	63.21
	Qwen3-VL-32B	66.19	14.85	60.14	81.85	22.80	56.44
Few-Shot	MedGemma-4B	43.63	7.91	36.51	53.90	29.44	**70.35**
	Rnj1-8B	35.09	**9.33**	29.98	59.92	18.51	62.64
	Qwen3-VL-8B	**53.47**	8.26	**46.09**	**73.31**	**32.53**	64.90
	OpenBioLLM-8B	10.18	0.95	8.32	41.73	3.92	4.55
	OLMo3-7B	29.32	5.77	26.74	47.31	28.33	62.30
	GPT-OSS-20B	61.88	18.30	65.18	71.70	40.94	70.03
	MedGemma-27B	62.38	13.47	55.03	75.68	38.53	67.53
	Qwen3-VL-32B	66.77	14.17	64.59	81.63	32.95	62.61
CoT	MedGemma-4B	**51.11**	8.75	42.44	64.22	35.04	70.21
	Rnj1-8B	37.34	8.41	35.86	62.52	37.88	68.25
	Qwen3-VL-8B	36.99	**10.54**	**54.24**	**77.23**	**46.94**	**70.72**
	OpenBioLLM-8B	9.74	1.18	17.92	38.85	38.49	38.34
	OLMo3-7B	36.64	7.52	33.52	56.67	35.84	63.55
	GPT-OSS-20B	26.89	14.78	62.02	84.54	28.51	52.49
	MedGemma-27B	**79.52**	**17.46**	**66.12**	78.66	47.44	72.42
	Qwen3-VL-32B	74.55	18.11	65.19	**84.95**	42.53	71.13

Table 4: Individual model accuracy (%) under three evaluation protocols across six benchmarks. **Bold** marks the best Panel 1 result per dataset per protocol. Highest marks the overall best per dataset. MXQA = MedXpertQA; MMed = MetaMedQA; PMed = PubMedQA; PMed+C = PubMedQA+C.

4B has the highest Panel 1 zero-shot accuracy on PubMedQA+C, suggesting benefit when biomedical context is provided. These observations describe patterns in this benchmark only and should not be read as universal claims about model capability.

CoT produces divergent effects: it lifts MedGemma-27B from 57.92% to 79.52% on MedQA ($p < 0.001$), establishing the strongest single-model result in our benchmark, while degrading OpenBioLLM-8B on multiple datasets. CoT prompting may conflict with some models' answer-format behavior, amplifying format errors rather than improving reasoning quality. Few-shot prompting severely degrades OpenBioLLM-8B on PubMedQA (41.11% zero-shot to 3.92% few-shot), consistent with negative transfer from examples whose format does not align with the model's response tendencies.

These results highlight that individual model performance is highly sensitive to prompting strategy and training alignment. A 22-point swing from CoT alone, and near-total failure under few-shot for one model, reinforces the need for evaluation frameworks that move beyond single-model accuracy to surface how models respond to external signals.

7 Panel-Based Evaluation

7.1 Overall CAMP Accuracy

Table 5 compares CAMP accuracy against the best individual model under each baseline protocol. The CAMP uses Panel 1 alone for 94.5 to 99.5% of questions, invoking Panel 2 only for 0.5 to 5.5% of escalated cases. CAMP outperforms the best individual Panel 1 zero-shot model on five of six datasets and remains close on MMLU-Med. Comparisons to CoT should be interpreted as comparisons across prompting settings rather than as direct zero-shot baselines.

The headline result is PubMedQA without context: CAMP achieves 54.07%, surpassing the best Panel 2 zero-shot model by 20.2 percentage points, the best zero-shot result by 12.96 points, and the best CoT result (MedGemma-27B, 47.44%) by 6.6 percentage points ($p < 0.001$). This result is notable because PubMedQA without context requires models to answer biomedical research questions

Dataset	CAMP Accuracy
MedQA	58.33
MedXpertQA	14.41
MetaMedQA	49.02
MMLU-Med	74.38
PubMedQA	54.07
PubMedQA+C	75.68

Table 5: CAMP accuracy across benchmarks. Individual model baselines are reported in Table 4.

Pathway	N	%	Acc.	Trigger
EC (Panel 1)	5,656	67.3	53.66%	AR ≥ 0.8
DC (Panel 1)	2,523	30.0	35.47%	1–3 rounds
P2C (Panel 2)	160	1.9	37.50%	P2 agree
WMV (fallback)	63	0.7	19.05%	P2 disagree

Table 6: CAMP resolution pathways for 8,402 questions with complete routing metadata. EC denotes Early Consensus, DC denotes Debate Consensus, P2C denotes Panel 2 Consensus, and WMV denotes Weighted Majority Vote. Percentages use the 8,402-question routing-metadata denominator.

without the source abstract, a regime where larger models might be expected to help. In this evaluated setting, CAMP's panel-based aggregation among five 4–8B models produces better results than any individual model and prompting setting we tested.

We hypothesize that panel aggregation can combine partially overlapping signals across models, enabling the collective elimination of incorrect options even when no single model has complete evidence. The three-class structure of PubMedQA (yes/no/maybe) may be amenable to this mechanism because the "maybe" class can reflect conflicting or insufficient evidence. This interpretation is suggestive rather than conclusive and requires broader validation.

7.2 Resolution Pathway Analysis

Table 6 reports the distribution and accuracy of CAMP resolution pathways for the 8,402 questions for which complete routing metadata were available. Percentages in this table are computed over this 8,402-question routing-metadata denominator.

Early consensus (EC) is the dominant pathway among questions with complete routing metadata, resolving 5,656 of 8,402 questions (67.3%), and is also the most accurate pathway (53.66%). The 18.19-point accuracy gap between EC (53.66%) and debate consensus (DC, 35.47%) should be interpreted primarily as difficulty triage rather than as evidence that debate itself is harmful: EC questions are those where Panel 1 models already show strong initial agreement, whereas DC questions enter debate precisely because Panel 1 is initially split. Within this harder contested group, debate provides a modest uplift of 4–5 percentage points over the Round 0 majority baseline.

Escalation pathways are rare in the routing-metadata analysis. Panel 2 consensus (P2C) accounts for 160 questions (1.9%), while the weighted-majority fallback is used for only 63 questions (0.7%). The low accuracy of WMV cases (19.05%) suggests that questions unresolved by both Panel 1 and Panel 2 may represent a collective knowledge boundary for the evaluated models. In a deployment-oriented setting, such fallback cases would be natural candidates for human review rather than automatic resolution.

8 Diagnostic Analysis

8.1 Agreement Ratio as a Difficulty Signal

One of the informative signals in the CAMP framework is the initial agreement ratio (AR, Equation 1). Figure 2 shows accuracy stratified by AR bin across all six datasets. For five of six datasets, the relationship between agreement, accuracy is generally positive. MedXpertQA is the exception: unanimous agreement (AR = 1.0) predicts only 6.6% accuracy, less than half the dataset mean of 14.4%. This inversion is associated with a systematic positional bias (Ko et al., 2020). Of 2,450 MedXpertQA, 22.8% of final answers select option A (558 total), while A is correct only 9.8% of the time (241 instances), a 2.3× over-selection ratio. This pattern may reflect shared training or evaluation artifacts, common instruction-tuning conventions, option-position bias, or correlated benchmark exposure; without training-data disclosure, we cannot isolate the source. We refer to the resulting behavior as correlated ignorance: not the stochastic uncertainty that deliberation can resolve, but a systematic shared error that debate may reinforce. The practical implication for benchmark is that agreement ratio is a useful dataset-level diagnostic. When a benchmark exhibits the normal positive AR-accuracy relationship, it can help calibrate escalation thresholds. When agreement and accuracy diverge, the benchmark exposes a qualitative failure boundary where deliberation may be counterproductive. The key takeaway is that agreement ratio is a reliable uncertainty proxy only under

Model	Stub.	LF-Idx	Total Inf.	Role
Qwen3-VL-8B	0.913	+0.306	4,274	Leader
OpenBioLLM-8B	0.892	+0.028	1,315	Outlier
OLMo3-7B	0.832	-0.006	1,480	Neutral
MedGemma-4B	0.806	+0.111	3,173	Active
Rnj1-8B	0.737	-0.012	2,233	Follower

Table 7: Per-model debate behaviour metrics averaged across all datasets. Stubbornness denotes the fraction of debate rounds in which a model maintains its position, LF-Idx denotes the leader–follower index from Equation 4, and Total Inf. reports cumulative outgoing influence events.

sufficient error independence (Kuai et al., 2026); when models share systematic biases, unanimity can signal failure rather than confidence.

8.2 Debate Dynamics and Model Roles

Position changes decay rapidly across debate rounds: Round 1 accounts for 67.7 to 81.8% of all position changes, Round 2 for 14.8 to 23.3%, and Round 3 for only 3.3 to 11.2%. This suggests that setting $R = 3$ rounds is adequate for the protocol studied here, although we do not claim optimality without ablations. The pattern also provides no evidence that simply increasing R would resolve MedXpertQA-class failures, where the dominant issue appears to be correlated error rather than unresolved disagreement. Table 7 characterises per-model debate behaviour. Qwen3-VL-8B has the highest outgoing influence count, exerting 4,274 cumulative influence events, 35% more than the next-highest model (MedGemma-4B, 3,173). The single strongest dyadic influence channel is Qwen3-VL-8B $\rightarrow$ Rnj1-8B (368 events on MedQA alone), consistent with Rnj1-8B's role as the panel's primary swing voter (lowest stubbornness: 0.737; most negative LF-index: -0.012). The combination of Rnj1-8B's flexibility and Qwen3-VL-8B's anchoring behaviour is associated with much of the panel's observed self-correction.

The paired PubMedQA experiment isolates the value of providing the source abstract. Accuracy rises from 54.07% on PubMedQA without context to 75.68% on PubMedQA+C, a 21.6 percentage point gain associated with context availability. Debate adds only 0.80 points on PubMedQA+C and 0.00 on PubMedQA without context. In this paired setting, access to the relevant evidence appears more important than adding further deliberation. For system builders, this suggests that retrieval or context provision should be evaluated before more elaborate debate protocols.

8.3 Error Decomposition

We categorise debate errors into two primary modes and analyze Panel 2 insufficiency separately for escalated cases. **Type 1 (debate insufficient):** the dominant failure (93 to 97% of all errors), where the panel converges on the wrong answer regardless of debate, often because models share the same misconception. **Type 2 (debate harmful):** a model initially correct was persuaded to switch away during debate (3 to 7% of errors). This corruption rate is relatively consistent across datasets, suggesting that harmful influence is not confined to a single benchmark. Separately, Panel 2 can remain insufficient after escalation, as on MedXpertQA where 76.87% of escalated cases remain wrong after Panel 2 processing. A cross-cutting finding is overconfidence on errors: models report 0.706 to 0.815 average confidence when they are wrong, approaching their confidence on correct answers. This confidence pattern limits the utility of confidence-weighted voting and suggests that confidence-gated escalation would require stronger calibration methods. Figure 3 visualizes the error-mode pattern and the wrong-answer confidence values.

8.4 Conservative Evidence on Design Choices

The current results provide limited, observational evidence about the CAMP design choices, but they should not be interpreted as full ablations. The 4-of-5 early-consensus threshold separates a high-agreement subset with higher accuracy from lower-agreement questions that require debate, while the MedXpertQA inversion shows why agreement must be treated as a diagnostic signal rather than a universal confidence guarantee. The observed decay in position changes across rounds supports the choice to cap debate at three rounds for this evaluation, but it does not prove that three rounds is optimal. Similarly, the low Panel 2 activation rate keeps the protocol compute-efficient, but the small escalated sample means conclusions about larger models must remain cautious.

Taken together, these findings support a conservative interpretation: CAMP is useful not because it always improves accuracy, but because it exposes when agreement, disagreement, escalation, and confidence behave differently across datasets. This diagnostic role is especially important in medical

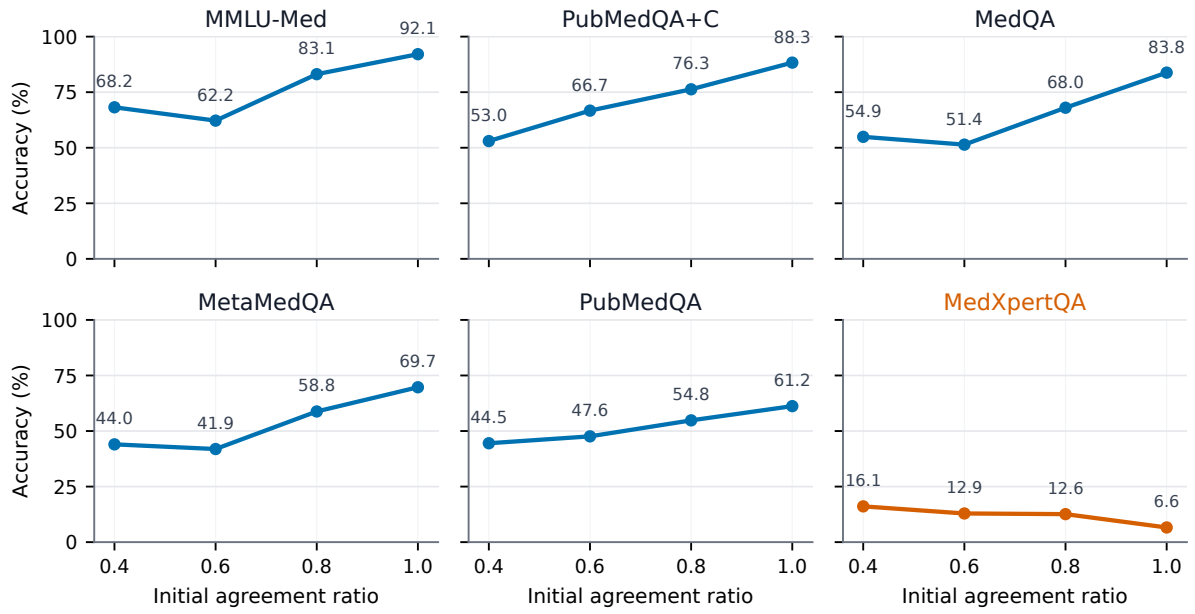


Figure 2: Accuracy conditioned on initial agreement-ratio bin. Five datasets show a generally positive relationship between agreement and accuracy, while MedXpertQA inverts this pattern, indicating correlated ignorance in this evaluation.

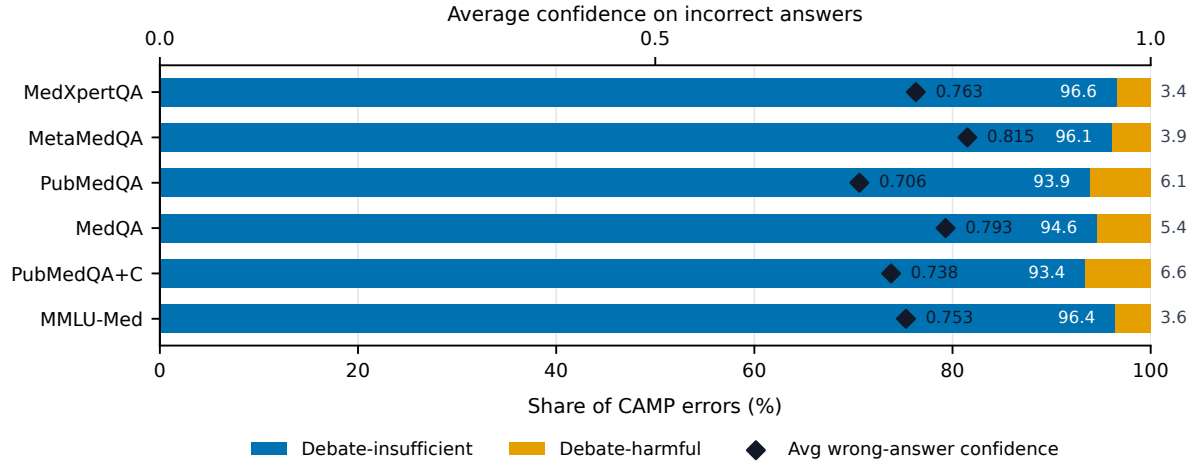


Figure 3: CAMP error decomposition across datasets. Debate-insufficient errors dominate on all datasets, while debate-harmful errors remain smaller but non-negligible; diamonds show the average confidence assigned to incorrect answers.

QA, where a high-accuracy aggregate can obscure whether the system is correcting errors, reinforcing shared mistakes, or deferring to larger models on too few cases to draw broad conclusions.

9 Conclusion

MedBench shows that evaluating medical language models only as isolated predictors misses important interaction-level behavior. Through CAMP, deliberation among smaller models improves performance in selected settings, particularly PubMedQA, but does not consistently replace scale or resolve expert-level gaps. More importantly, the protocol exposes when consensus is informative and when it is misleading: agreement tracks difficulty across most datasets, while MedXpertQA reveals correlated ignorance. The observed debate-insufficient failures, harmful persuasion, and over-confidence on wrong answers suggest that deliberative evaluation should complement accuracy-based benchmarking, with future work on panel design, calibration, model diversity, and human oversight.

Limitations

Knowledge ceiling: MedXpertQA demonstrates that deliberation does not reliably overcome expert-level gaps for the evaluated 4B to 8B models. At 14.4% accuracy (barely above 10% random), this dataset marks a difficult boundary for the current panel. Our results do not show evidence that simply increasing debate rounds would resolve this failure mode.

Correlated training data: Panel 1 models may share pretraining, instruction-tuning, evaluation, or formatting artifacts, which could amplify correlated failure modes such as the positional bias observed on MedXpertQA. Because training-data disclosure is incomplete, we cannot isolate the source of this correlation. A panel with more diverse knowledge sources might exhibit different deliberation dynamics.

Small Panel 2 sample sizes: Panel 2 handles only 0.5 to 5.5% of questions per dataset (4 to 134 questions). Panel 2 statistics (particularly for PubMedQA where only 6 questions escalated) should be interpreted cautiously.

No ablations on panel design: We did not ablate panel size (3 vs. 5 models), debate round limit (1 vs. 2 vs. 3), or escalation threshold. These design choices likely affect both accuracy and the diagnostic value of agreement ratio as a signal.

Confidence calibration: The observed overconfidence on incorrect answers prevents confidence from serving as a useful escalation gate in this study. Calibration-aware prompting or post-hoc calibration strategies are needed before confidence can reliably inform routing decisions.

Ethical Considerations

Clinical deployment risk: MedBench is an evaluation benchmark; the CAMP is an evaluation protocol. The 41.7% error rate on MedQA and 85.6% on MedXpertQA make clear that neither the individual models nor the panel are ready for autonomous clinical deployment without extensive validation, regulatory approval, and mandatory human oversight.

Dataset biases: MedQA and MMLU-Med reflect US-centric medical education. MetaMedQA provides some multilingual coverage, but demographic and geographic biases in underlying training data may produce unequal performance across patient populations.

Overconfidence risk: Models report high average confidence (0.706 to 0.815) on incorrect answers. Any downstream clinical decision support tool must prominently communicate this overconfidence risk to end users.

Acknowledgements

We gratefully acknowledge the Koita Centre for Digital Health, IIT Bombay, and IIT Bombay Trust Lab for their support, resources, and funding that enabled this work. The authors received funding for this project from IIT Bombay Trust Lab. We are sincerely grateful for their generous support. We also thank Mayank Kumar for participating in helpful discussions.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Pedro Cisneros-Velarde. 2024. On the principles behind opinion dynamics in multi-agent systems of large language models. *Preprint*, arXiv:2406.15492.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2567–2577. ArXiv:1909.06146.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. *Preprint*, arXiv:2004.14602. ArXiv:2004.14602.

Chenchen Kuai, Jiwan Jiang, Zihao Zhu, Hao Wang, Keshu Wu, Zihao Li, Yunlong Zhang, Chenxi Liu, Zhengzhong Tu, Zhiwen Fan, and Yang Zhou. 2026. How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles. *Preprint*, arXiv:2604.07650. ArXiv:2604.07650.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36. ArXiv:2303.17760.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2305.19118.

Ehsan Misaghi, Sean T Berkowitz, Bing Yu Chen, Qingyu Chen, Renaud Duval, Pearse A Keane, Danny A Mammo, Ariel Yuhan Ong, Mertcan Sevgi, Sumit Sharma, Sunil K Srivastava, Yih Chung Tham, and Fares Antaki. 2026. Deliberative multi-agent large language models improve clinical reasoning in ophthalmology. *Preprint*, arXiv:2603.21447.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *Preprint*, arXiv:2303.13375. ArXiv:2303.13375.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenBioLLM. 2024. Llama3-OpenBioLLM-8B model card. Hugging Face model card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. ArXiv:2203.14371.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Ashish Vaswani, Mike Callahan, Adarsh Chaluvaraju, Aleksa Gordić, Devaansh Gupta, Yash Jain, Divya Mansingka, Philip Monk, Khoi Nguyen, Mohit Parmar, Michael Pust, Tim Romanski, Peter Rushton, Ali Shehper, Divya Shivaprasad, Somanshu Singla, Kurt Smith, Saurabh Srivastava, Anil Thomas, and 4 others. 2025. Rnj-1-Instruct. Instruction-tuned model release.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*. ArXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35. ArXiv:2201.11903.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations*. ArXiv:2306.13063.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the 42nd International Conference on Machine Learning*. ArXiv:2501.18362.

Fast, Accurate, and Local Conversion of MIMIC-IV to OMOP with DBT

Adam Sutton[1], Niko Möller-Grell[1], Thomas Searle[1], Richard Dobson[1,2,3]
[1] Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience King's College London, London, UK
[2] Institute of Health Informatics, University College London, London, United Kingdom
[3] NIHR Mausley Biomedical Research Centre, London, United Kingdom
Correspondence: adam.sutton@kcl.ac.uk

Abstract

dbt_mimic_omop is a free, open-source resource that converts the MIMIC-IV dataset to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) format on consumer level hardware. CDM approaches are increasingly adopted in both industry and academia due to the need for interoperability and reproducibility, including in clinical NLP tasks such as cohort selection, information extraction, and retrieval-augmented generation. The MIMIC-IV database is among the most widely used critical care research datasets, yet existing pipelines to transform it to OMOP depend on enterprise database infrastructure and complex orchestration, limiting accessibility for practitioners and resource-constrained researchers. We further integrate free-text clinical notes (195.6M clinical annotations) and chest radiographs into the OMOP note_nlp and imaging extension tables, making all MIMIC-IV modalities (structured data, free-text, and imaging) accessible through a common data model. This resource generates a more comprehensive dataset than existing alternatives and is intended to be used to aid in system development, testing, and evaluation.

1 Introduction

A growing number of clinical datasets are becoming available to the research community (All of Us Research Program Investigators, 2019; Clark et al., 2013; Pollard et al., 2018; Johnson et al., 2023b). However, each resource adopts its own bespoke schema, limiting interoperability and increasing development overhead of multi-source studies.

Common Data Models (CDMs) address this by enabling consistent data representation across institutions and applications through a standardised framework of schemas, attributes, entities, and relationships, thus facilitating collaboration, benchmarking, and reproducible analytics (Finster et al., 2025; Brown et al., 2022; PCORnet, 2025; Garza et al., 2016).

These data models have proven to be useful in emerging fields within healthcare and computational linguistics, such as information extraction and retrieval augmented generation (Keloth et al., 2023; Möller-Grell et al., 2026).

The Medical Information Mart for Intensive Care (MIMIC-IV) database is a cornerstone of critical care research, providing granular clinical data for thousands of admissions to the ICU (Johnson et al., 2023c). The MIMIC-IV data collection covers a wide range of data modalities, from structured data and textual data to images.

However, the joint analysis of these disparate data sources can be resource intensive, requires custom adaptations, and is not easily extensible to additional data sources. Converting these datasets into a CDM such as the OMOP CDM provides a unified extensible representation to cross-institutional data and a consistent foundation for systematic annotation of clinical text via Natural Language Processing (NLP).(Observational Health Data Sciences and Informatics, 2026).

Existing MIMIC ETL pipelines rely on enterprise database infrastructure (BigQuery) and complex orchestration (Paris et al., 2021; Sciences and Informatics, 2020), creating barriers for those who are resource-constrained.

dbt_mimic_omop' [1] is a lightweight, modular and portable ETL pipeline that uses dbt-Core (data build tool) ((dbt Labs, Inc., 2026)) and DuckDB ((Raasveldt and Mühleisen, 2019)) to transform MIMIC-IV data into the OMOP CDM (v5.4).

In addition to the core MIMIC-IV dataset, we also offer conversion for supplementary MIMIC-IV datasets. MIMIC-IV-Note contains more than 2 million de-identified free text notes, MIMIC-IV-CXR is an imaging database with more than

[1] link removed for reivew

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 992–996
July 3-4, 2026 ©2026 Association for Computational Linguistics

300,000 medical images. Both datasets are linked to the same patients as the core dataset. Further to this, we provide a resource with clinical annotations for the free text data in MIMIC-IV-Note.

2 Data Sources

2.1 MIMIC-IV

MIMIC-IV is well known in multiple fields of research, providing access to de-identified healthcare data. MIMIC-IV contains data for more than 65,000 patients admitted to the ICU and more than 200,000 patients admitted to the emergency department (Johnson et al., 2023a). The dataset does not adhere to any common data model.

2.2 MIMIC-IV-Note

MIMIC-IV-Note contains 331,794 de-identified discharge summaries from 145,915 patients and 2,321,355 radiology reports and for 237,427 patients, from the same sample of patients from the core MIMIC-IV dataset.

2.3 MIMIC-IV-CXR

The MIMIC Chest X-ray Database (MIMIC-CXR) is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports (Johnson et al., 2019). The dataset contains 377,110 images corresponding to 227,835 radiographic studies. This dataset uses the CheXpert labelling system (Irvin et al., 2019), 14 distinct radiographic diagnostic observations through rule-based extraction are mapped to SNOMED concept IDs via custom vocabulary mappings.

2.4 CogStack Dashboards

CogStack Dashboards is an OpenSearch cluster that hosts the MIMIC-IV and MIMIC-IV-Note dataset[2]. In addition to the MIMIC dataset, annotations have been generated, provided via MedCAT (Kraljevic et al., 2021; Sutton et al., 2026). These generated annotations use the SNOMED CT ontology (SNOMED International, 2025) to provide coding for all detected entities. They also provide additional information about the entity, such as meta information and character positioning within the text. This model has been trained on various open-source datasets as well as privately trained data from King's College Hospital and Guy's and St Thomas' Hospital. Annotations such as these are useful in information extraction and retrieval augmented generation tasks (Keloth et al., 2023; Möller-Grell et al., 2026).

[2]https://cogstackdashboards.sites.er.kcl.ac.uk/

2.5 OHDSI Athena

OHDSI Athena is the repository for OHDSI standardised vocabularies used in the OMOP CDM. It allows users to search, browse, and download standardised mapped medical codes (concepts) for clinical, drug, and procedure data, allowing harmonisation and interoperability (Observational Health Data Sciences and Informatics, 2025).

3 Implementation

The DBT project's medallion architecture to transform raw MIMIC-IV to OMOP is as follows:

- **Bronze Layer (Ingestion):** Raw MIMIC-IV CSV files (v3.1), MIMIC-IV-CXR metadata and radiology reports, and OHDSI Athena vocabularies (v5) are ingested into a local DuckDB database. Python scripts facilitate efficient loading of large compressed datasets directly from Parquet/CSV sources. For MIMIC-IV-CXR, DICOM metadata is extracted and loaded alongside associated free-text radiology reports, which are linked to the patient and study identifiers shared with the core MIMIC-IV dataset.
- **Silver Layer (Refinement):** Intermediate views handle data cleaning, type casting, and semantic mapping. Specific transformations include cleaning visit details from various source tables (such as admissions/transfers) and mapping source codes (ICD-9/10, NDC, LOINC, local microbiology codes) to standard OMOP concepts using the OHDSI Athena vocabulary lookup tables. CXR radiology reports are linked to their corresponding MIMIC-IV hospital admission and processed for downstream annotation. CXR studies are linked to OMOP visit occurrence records by temporal overlap, matching subject ids and cross-referencing visit windows.
- **Gold Layer (Standardisation):** The final OMOP CDM v5.4 tables are compiled, covering key OMOP tables such as person, visit occurrence, condition occurrence, drug exposure, measurement, procedure occurrence, and observation. Tables generated from supplementary data do not contribute to the core

Domain	Field	PostgreSQL	dbt/DuckDB
condition_occurrence	condition_concept_id	100%	100%
procedure_occurrence	procedure_concept_id	100%	100%
drug_exposure	drug_concept_id	86.10%	96.59%
drug_exposure	route_concept_id	0%	97.79%
measurement	measurement_concept_id	0%	95.13%
measurement	unit_concept_id	0%	98.60%
specimen	specimen_concept_id	0%	97.53%
person	gender_concept_id	100%	100%
person	race_concept_id	0%	52.23%
person	ethnicity_concept_id	0%	0% (unmapped)

Table 1: Key OMOP Concept Coverages - each percentage represents the number of rows in each corresponding table that has appropriate concept ID in the specified field. These fields are required by the OMOP CDM standard.

Table	Count
measurement	296705150
note_nlp	**195642167**
fact_relationship	74356046
drug_exposure	20351014
observation	14018411
drug_era	10296936
procedure_occurrence	6564445
condition_occurrence	5328942
note	**2652887**
image_feature	**656697**
image_occurrence	**372209**
person	364627
death	38301

Table 2: Key OMOP Table Counts

OMOP CDM tables. MIMIC-IV-Note and annotations from CogstackDashboards are inserted into the note and note_nlp tables respectively. Tables sourced from MIMIC-CXR are generated into the OMOP CDM MI working group extension tables: image_occurrence with visit linkage, view position, modality, local and google cloud file paths; and image_feature, with one row per CheXpert finding, certainty labels mapped per custom SNOMED mapping.

The modular dbt project structure separates logic into 69 distinct models at transformation stage (e.g., separate models for admissions, pharmacy, labevents), facilitating maintenance and community contributions.

4 Results

The pipeline transforms MIMIC-IV v3.1 to OMOP CDM v5.4, generating more than 343 million standardised clinical events in 364,627 patients. On a consumer laptop (16gb of RAM, 8 core cpu), the full ETL process can be completed in less than three hours.

Tab. 2 shows the record counts of key OMOP tables and the tables generated from the supplementary data source. Tab. 1 shows the mapping rates for key concepts within this implementation.

Our proposed pipeline recovers previously unmapped domains, most notably measurement concept ids ($0\% \rightarrow 95.13\%$) and specimen concept ids ($0\% \rightarrow 96.59\%$), and introduces full coverage for all imaging fields. Race concepts improve from 0% to 52.23%, while ethnicity concept ids remains unmapped at 0% in both implementations. These lesser performances are due to the vocabularies lacking standard translations from MIMIC-IV recording of ethnicity and race to OMOP standardized codes.

We also directly compare the coverage between our solution and Legrand et al. (2025). Tab. 3 shows raw record counts and concept mapping rates for both solutions.

Free text clinical notes are also included, sourced from MIMIC-IV-Note. These notes are imported into the note table. The pipeline also incorporates full NLP annotation processing via the note_nlp table, adding 195.6M structured annotations from previously unprocessed clinical notes—a capability absent from prior implementations. However, due to the significant processing cost, we also offer the table already transformed ‘note_nlp‘ as a download,

Table 3: Comparison of PostgreSQL solution (Legrand et al., 2025) vs dbt/DuckDB Metrics

Metric	PostgreSQL	dbt/DuckDB	Change
Record Counts (millions)			
Person	0.32M	0.36M	+14.6%
Measurement	169.1M	296.7M	+75.4%
Procedure	0.86M	6.56M	+7.6×
Observation	3.52M	14.02M	+4.0×
Concept Mapping Rates			
Condition concepts	100%	100%	—
Drug concepts	86.1%	96.6%	+10.5%
Measurement concepts	0	95.1%	+95.1%
Unit concepts	0%	98.6%	New

with instructions provided in the repository.

In addition to the core OMOP CDM tables, the pipeline populates the OMOP Medical Imaging (MI) working group extension tables. The table image_occurrence contains 372,209 chest X-ray images, and image_feature captures 656,697 CheXpert findings with 100% concept mapping across all imaging fields (modality, anatomic site, image type concept and feature type). Of these images, 330,401 (88.8%) were successfully linked to visit_occurrence via temporal mapping.

5 Conclusion

The ‘dbt_mimic_omop‘ pipeline is freely available on GitHub [3]. This open-source solution significantly reduces the technical barrier for researchers wishing to conduct OHDSI-compliant studies using MIMIC-IV data. By decoupling the ETL process from enterprise-grade infrastructure, we facilitate broader adoption in educational settings, reproducibility in research, and rapid prototyping of analytical pipelines.

This resource is intended to be used as an aid for those who are developing NLP systems that employ the OMOP Common Data Model. The free text that is available via ‘mimic-iv-note‘ is available and in OMOP format for various tasks such as entity recognition and linking. The generated annotations we have provided can be used as baseline comparison. The dataset could also find use as a test dataset for more complex pipelines such as Retrieval Augmented Generation.

By combining dbt’s software engineering best practices with DuckDB’s embedded analytical capabilities, and extending the CDM with note_nlp and imaging tables, we aim to lower the barrier to clinical NLP, multimodal system development, and reproducible observational research over MIMIC-IV.

Future work includes improvements to this ETL pipeline. Such as additional versions of the MIMIC datasets (Johnson et al., 2016), and other publicly available datasets. Future versions of the repository can also include additional supplementary datasets (Moody et al., 2022). Such pipelines can also speed up the benchmarking process. Various datasets and benchmarking challenges are available, but often require development time for that specific challenge (DrivenData, 2026). Work can be done making such challenges follow a common data model, if applicable.

[3] https://github.com/CogStack/dbt_mimic_omop

References

All of Us Research Program Investigators. 2019. The “all of us” research program. *New England Journal of Medicine*, 381(7):668–676.

Jeffrey S Brown, Aaron B Mendelsohn, Young Hee Nam, and 1 others. 2022. The us food and drug administration sentinel system: a national resource for a learning health system. *Journal of the American Medical Informatics Association*, 29(12):2191–2200.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. 2013. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057.

dbt Labs, Inc. 2026. dbt (Data Build Tool). Technical report, dbt Labs, Inc. Accessed: 2026-02-20.

DrivenData. 2026. Snomed ct entity linking benchmark. https://www.drivendata.org/benchmarks/

310/benchmark-snomed-ct/. Accessed: 2026-02-25.

Melissa Finster, Markus Wenzel, and Elham Taghizadeh. 2025. Common data models and data standards for tabular health data: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1):422.

Maryam Garza, Guilherme Del Fiol, Jessica Tenenbaum, Anita Walden, and Meredith Nahm Zozus. 2016. Evaluating common data models for use with a longitudinal community registry. *Journal of biomedical informatics*, 64:333–341.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 590–597.

Alistair Johnson, Tom Pollard, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023c. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(160035).

Vipina K Keloth, Juan M Banda, Michael Gurley, Paul M Heider, Georgina Kennedy, Hongfang Liu, Feifan Liu, Timothy Miller, Karthik Natarajan, Olga V Patterson, and 1 others. 2023. Representing and utilizing clinical textual data for real world studies: An ohdsi approach. *Journal of Biomedical Informatics*, 142:104343.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Paul Legrand, Kawsar Noor, Satyam Bhagwanani, and Richard Dobson. 2025. An open-source text-to-sql pipeline for omop-formatted electronic health records. Technical report, King's College London. Available from https://github.com/red1dwarf/text-to-omop/tree/main/ETL (accessed 2026-02-20).

Niko Möller-Grell, Linglong Qian, and Shihao Shenzhang. 2026. Multimodal ards detection through agentic medgemma workflows. https://www.kaggle.com/competitions/med-gemma-impact-challenge/writeups/new-writeup-1770377682785. Kaggle MedGemma Impact Challenge write-up. Accessed: 2026-02-25.

Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. 2022. MIMIC-IV Waveform Database. *PhysioNet*. Version 0.1.0.

Observational Health Data Sciences and Informatics. 2025. ATHENA – OHDSI standardized vocabularies.

Observational Health Data Sciences and Informatics. 2026. OMOP Common Data Model. Technical report, OHDSI. Accessed 2026-02-20.

Nicolas Paris, Antoine Lamer, and Adrien Parrot. 2021. Transformation and evaluation of the mimic database in the omop common data model: development and usability study. *JMIR Medical Informatics*, 9(12):e30970.

PCORnet. 2025. Pcornet common data model v7.0 specification. Technical report, Patient-Centered Outcomes Research Network (PCORnet).

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178.

Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 international conference on management of data*, pages 1981–1984.

Observational Health Data Sciences and Informatics. 2020. Ohdsi/mimic: Etl for transforming mimic-iv to the omop common data model. Technical report, OHDSI. GitHub repository. Accessed 2026-02-20.

SNOMED International. 2025. SNOMED CT. International Health Terminology Standards Development Organisation (IHTSDO) / SNOMED International, International Edition.

Adam Sutton, Vlad Dinu, Thomas Searle, and Richard Dobson. 2026. Cogstack dashboards. Accessed: 2026-02-20.

Exploring Novel Drug Research Area using Large Language Models Based on Research Trends in Biomedical Literature

Afnan[1], Michael Van Supranes[1,2], Tomohiro Nishiyama[1], Shoko Wakamiya[1], Eiji Aramaki[1]

[1]Nara Institute of Science and Technology, Japan, [2]University of the Philippines, Philippines

Abstract

The rapid expansion of biomedical literature makes manual identification of novel drug-disease relationships increasingly difficult. Existing approaches have leveraged LLMs to mine abstracts or construct knowledge graphs for drug repurposing. There are two key limitations: finite context windows for capturing macro-level research trends, and single-pass black-box pipelines make it difficult to verify outputs. This paper proposes a pipeline for discovering new drug targets by combining disease and drug research trends using Large Language Models (LLMs). Our method extracts PICO components from PubMed abstracts, normalizing the Population and Intervention Component to ICD and ATC codes, respectively. A *temporal frequency delta* matrix is constructed to capture publication count shifts across 2013 to 2022, then used to discover novel drug areas. Compared with the abstract-based baseline, our approach showed qualitative signs of generating combinations that were more closely aligned with observed research trends and, in some cases, more clinically plausible. These findings suggest the potential usefulness of structured trend information for LLM-based exploration, although the differences between the two methods were limited and the results remain preliminary. Future work will focus on validating the consistency and reliability of these candidates.

Keywords: LLMs, drug re-purposing, biomedical NLP, research trend, PubMed

1 Introduction

The biomedical literature has grown exponentially with time, creating a rich and complex body of knowledge for scientific discovery. Although experienced researchers can generate novel ideas from prior work, there are inherent limits to the amount of information that can be effectively processed and compared within a given time. As a result, identifying relevant studies and distilling key insights that can lead to breakthroughs has become increasingly challenging amid the vast and rapidly expanding volume of publications.

This challenge is pronounced in domains such as oncology, where thousands of abstracts are published each year, advancing research on drug-disease interactions, clinical trial outcomes, and molecular pathways (Wang et al., 2025; Yan et al., 2024). In this context, large language models (LLMs) have emerged as powerful tools for assisting researchers in navigating large-scale literature. LLMs can support automatic synthesis of clinical evidence and uncover hidden knowledge connections that would otherwise take a long time and require extensive manual review (Ghafarollahi and Buehler, 2025). In addition to literature review, recent innovations have demonstrated that LLMs can also assist in generating novel, plausible clinical hypotheses, such as identifying drug re-purposing candidates for Alzheimer's disease and Acute Myeloid Leukemia (Yan et al., 2024; Gottweis et al., 2025).

Despite the potential of LLM-based scientific discovery, direct application of LLMs to large-scale literature analysis has the following limitations:

Limited ability to capture overall research trends: LLM-based methods are constrained by a finite context window, limiting the number of abstracts or related work that can be included in a single prompt. LLMs may also fail to fully integrate all important findings when the input is too long. This phenomenon is called *lost in the middle* or retrieval failure during long-context processing (Du et al., 2025; Li et al., 2025). Thus, using LLMs requires selective inclusion of related work, which may hinder the model's ability to capture broader research trends.

Limited transparency in LLM-based pipelines: LLMs are black boxes and often deployed in single-pass pipelines for complex tasks. As a result, it is

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 997–1013
July 3-4, 2026 ©2026 Association for Computational Linguistics

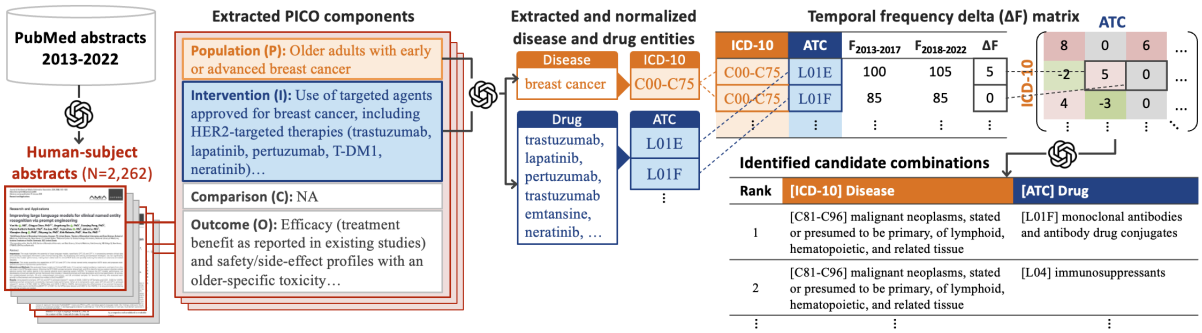


Figure 1: Analysis pipeline of the study. Abstracts were first filtered to include only human populations, then PICO components (Population, Intervention, Comparison, and Outcome) were extracted. The Population (P) and Intervention (I) components were then used to construct a matrix of publications for each P–I combination across two periods: 2013–2017 and 2018–2022. This matrix served as input for an LLM to generate novel potential P–I combinations. Multiple reasoning and an impact score in each step of generating are used to mitigate hallucination.

difficult to tell whether the model has correctly identified and incorporated relevant information or has generated hallucinated content (Wang et al., 2025).

Addressing these challenges remains an open problem. In this paper, we present a novel approach to support drug discovery by augmenting LLMs with temporally grounded bibliometric signals (Figure 1). Our contributions are two-fold. First, to address the difficulty LLMs face in detecting macro-level research trends across fields and time periods, we introduce a frequency delta matrix to represent research trends in a compact, context-efficient manner, enabling the incorporation of research momentum signals into prompts without requiring extensive collections of raw abstracts. Second, we develop an end-to-end pipeline that extracts structured information from abstracts based on the PICO framework (Population (**P**), Intervention (**I**), Comparison or Control (**C**), and Outcome (**O**)), constructs co-occurrence matrices, and leverages them to guide LLMs in finding candidate disease–drug research combinations. We compare this structured, trend-explicit formulation with a baseline that uses raw abstracts and publication years as unstructured, trend-implicit input, and examine whether LLMs can improve in generating more clinically plausible disease–drug research combinations.

2 Related Work

2.1 LLMs in Biomedical Discovery

The integration of LLMs into the medical field has shown immense promise, particularly in automating knowledge synthesis. Early benchmarks, such as BioRED (Du et al., 2025), established a gold standard for extracting explicit semantic triples (e.g., Drug A treats Disease B) from unstructured PubMed abstracts. Beyond simple extraction, LLMs are now being utilized as autonomous agents capable of generating novel and scientifically valid research ideas (Baek et al., 2025; Gottweis et al., 2025). However, a persistent challenge remains in ensuring the interpretability and reliability of these discoveries, especially when models move from sentence-level connections to complex, multi-faceted scientific problem-solving (Li et al., 2025; Williams, 2026).

2.2 Automated Hypothesis Generation in Drug Discovery

The primary objective of AI-driven drug discovery is to identify novel combinations of diseases and pharmacological agents that have not yet been clinically explored. Traditional methods relied on building discrete concept corpora and identifying relationships via word vector similarities (Islamaj et al., 2024) or link prediction within knowledge graphs (Yang et al., 2023). While these methods are effective for validating known biochemical pathways, they often fail to capture emerging research frontiers or trending therapeutic dyads that exhibit high growth momentum. Recent advances use the scientific knowledge in LLMs to move beyond predicting localized connections and toward identifying "hot topics" in science, a task traditionally handled by bibliometric analysis (Qi et al., 2023; Tshitoyan et al., 2019). Our work builds on this by investigating whether LLMs can automate the interpretive step of translating statistical clusters into actionable research hypotheses.

2.3 Topic Trends as Scientific Signals

The use of publication frequency and temporal shifts to identify emerging research directions is a well-established principle in scientometrics. Traditional bibliometric methods, such as co-word analysis and citation burst detection, have long been used to map the "intellectual turning points" of a scientific field (Qi et al., 2024). For instance, a sudden spike in publications on a specific drug-disease dyad often signals a transition from basic research to clinical interest. Our work builds on the premise that scientific discovery is a temporal process, in which "information remnants" in literature, captured via frequency shifts, can predict future high-impact research areas (Tshitoyan et al., 2019). By transforming bibliometric data into a *frequency delta matrix* (Δ), we construct a compact representation of research trends that can be efficiently incorporated into LLM prompts, providing signals of research momentum that are difficult to capture when using raw abstracts alone.

2.4 Reasoning over Tabular Data

While LLMs are natively trained on unstructured natural language, their ability to interpret structured formats like CSV, JSON, and SQL has become a focal point of research. Recent studies indicate that while LLMs can perform reasoning over tabular data, their performance is highly sensitive to the data's serialization format (Sybrandt et al., 2017). In the context of drug discovery, the ability of a model to ingest structured co-occurrence matrices is essential for identifying statistical trends. Researchers have found that providing LLMs with structured summaries of biomedical metadata enables more precise "*chain-of-thought*" reasoning than processing raw, disconnected abstracts (Kim et al., 2022). This study extends this line of inquiry by comparing structured tabular inputs.

3 Dataset Construction

The primary dataset for this study comprises scholarly abstracts retrieved from PubMed, spanning a 10-year period from 2013 to 2022. This period was selected to enable a balanced comparison between two consecutive five-year intervals (2013–2017 and 2018–2022). Using equal time intervals helps ensure that observed differences reflect meaningful shifts in research activity. In addition, this period captures the accelerated growth in oncology research.

Year	# of filtered articles	Avg. word count
2013	128	251
2014	169	242
2015	152	245
2016	176	245
2017	171	251
2018	168	249
2019	172	252
2020	293	250
2021	433	251
2022	400	177
Total	**2262**	–
Min	**128**	–
Max	**400**	–

Table 1: Number of abstracts after filtering (not animal or biomedical abstracts, only clinical abstracts), along with average word counts per year.

To construct a high-quality dataset for analysis, it is essential to accurately identify clinically relevant studies. However, traditional keyword-based Boolean searches often suffer from low precision due to polysemy and the inclusion of non-human studies.

The dataset was constructed in a two-stage process. First, an initial pool of 5,000 abstracts was retrieved from PubMed[1] via the API[2], with 500 abstracts randomly sampled per year to balance analytical coverage and computational feasibility. Second, each abstract was screened using an LLM (GPT-5, OpenAI[3]) to retain only those describing human studies (see Section 4.1 for details). This process yielded 2,262 abstracts, which constitute the final dataset used in this study.

Further details on the dataset distribution are provided in Table 1, which shows that annual publications remained below 180 between 2013 and 2018, exceeded 200 by 2020, and nearly doubled by 2021–2022. These trends have been attributed to advances in genomics and immunotherapy, increased clinical trial activity, and the growing use of artificial intelligence in data analysis (Yao et al., 2023; Wang et al., 2024; Azmi and Howe Chan, 2025).

4 Method

The overall methodology of this study consists of four main stages: LLM-based abstract screening and PICO extraction (Section 4.1), entity extraction

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://www.ncbi.nlm.nih.gov/books/NBK25497/

[3] https://developers.openai.com/api/docs/models/gpt-5

and normalization (Section 4.2), construction of the temporal frequency delta matrix (Section 4.3), and identification of potential disease-drug combinations from trending topics (Section 4.4). The complete pipeline is illustrated in Figure 1.

4.1 PICO Extraction

We use an LLM-based approach to identify the Population, Intervention, Comparison, and Outcome (PICO) components in abstracts. For each abstract, the LLM first determines whether the study concerns a human population; only if this criterion is satisfied does it proceed to extract the PICO components. This design integrates dataset screening with information extraction in a single process.

To ensure high accuracy and mitigate potential hallucinations, we employ a one-shot prompting strategy with multi-step reasoning. Before producing the final output, the LLM is also instructed to generate intermediate reasoning steps. This approach helps ensure that the extracted PICO components are grounded in the terminology used in each abstract. The detailed prompt provided in Appendix A and the example of the output is provided in Appendix G.

4.2 Extraction and Normalization of Disease and Drug Entities

Based on the extracted PICO components, we identify structured biomedical entities for downstream analysis. The `P` component identifies the underlying pathology or patient condition, while the `I` component captures the therapeutic agents evaluated in the study.

To ensure standardization and interoperability, we use an LLM-based approach for entity extraction and normalization. In this approach, the same LLM is used for both entity extraction and ontology mapping. First, we extract all the disease and drug entities (therapeutic agents) mentioned in the `P` and `I` components. These entities are then mapped to established biomedical ontologies: disease entities to the International Classification of Diseases (ICD-10, version 2019) and drug entities to the Anatomical Therapeutic Chemical (ATC) classification.

Both ontologies provide hierarchical structures, enabling each entity to be represented at multiple levels of granularity, with deeper levels corresponding to more specific clinical concepts. In the mapping process, we begin at the most specific category level. If no match is found at that level, the LLM escalates to the next higher category, continuing upward until a suitable match is identified. Regardless of the level at which the match is found, all hierarchical levels above it are retained, preserving the full ontological path of each entity. To balance clinical specificity with statistical robustness, we standardize the granularity of our variables, as detailed in Figures 3 (ICD-10) and 4 (ATC) in Appendix H. This approach enables a high-resolution view of therapeutic trends while avoiding over-segmentation or non-representative categories.

This LLM-driven matching process enables flexible semantic normalization of noisy biomedical text, including synonym resolution and implicit concept alignment, without relying on explicit string matching or rule-based heuristics. For further information, refer to Appendix B.

4.3 Temporal Frequency Delta Matrix

To investigate the evolution of research priorities, we divide the dataset into two distinct five-year cohorts: an early period (2013–2017) and a recent period (2018–2022).

By segmenting the data into these time windows, we quantify shifts in scientific interest over a decade and identify therapeutic interventions that have gained or declined in prominence.

We then construct a matrix to capture the frequency of disease-drug combinations within each period. Based on these matrices, we define the primary metric for our heatmap, the *Temporal Frequency Delta* (ΔF), which represents the net change in publication volume for a given pairing between the two cohorts.

$$\Delta F(P, I) = \sum_{t=C+T}^{C+2T-1} F_t(P, I) - \sum_{t=C}^{C+T-1} F_t(P, I) \quad (1)$$

where C denotes the starting year (set to 2013 in this study), and T denotes the length of each time window (set to 5 years). $F_t(P, I)$ denotes the number of unique publications discussing a specific disease–drug pair in year t, where P and I denote the population (disease) and intervention (drug), respectively. This formulation captures the temporal trend as the difference in publication counts between the two periods (2013-2017 and 2018-2022).

The ΔF value can be positive, negative, or zero:

- **Positive** ΔF indicates trending therapeutic combinations, where research activity has increased in the recent period. These regions are the primary focus of our LLM-based analysis.
- **Negative** ΔF indicates combinations with declining research activity over time.

- **Zero** ΔF indicates no change in the number of publications for a given disease–drug pairing between the two periods.

Finally, using the combinations of P and I, we define the *temporal frequency delta* matrix, $\Delta F \in \mathbb{R}^{n \times m}$, as follows:

$$\begin{pmatrix} \Delta F(P_1, I_1) & \Delta F(P_1, I_2) & \cdots & \Delta F(P_1, I_m) \\ \Delta F(P_2, I_1) & \Delta F(P_2, I_2) & \cdots & \Delta F(P_2, I_m) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta F(P_n, I_1) & \Delta F(P_n, I_2) & \cdots & \Delta F(P_n, I_m) \end{pmatrix} \quad (2)$$

The lists of all $P_{1..n}$ and all $I_{1..m}$ are provided in Appendix E and F. In this study, $n = 51$ represents the number of diseases analyzed (focused on Neoplasm). While $m = 256$ denotes the total number of drugs analyzed.

4.4 Identification of Candidate Disease–Drug Combinations from Emerging Trends

To identify candidate combinations for future research, the proposed method uses a structured, matrix-based representation derived from temporal publication signals. This design enables a controlled evaluation of how input formulation influences an LLM's ability to synthesize meaningful and novel research areas. The method is provided with tabular data consisting of P–I pairs and *temporal frequency deltas* (ΔF), which capture shifts in research priorities.

The LLM is instructed to select the top-k ($k = 10$ in this study) most promising P–I combinations based on four scoring criteria: growth over the past five years, translational plausibility, clinical relevance, and strategic novelty. Each criterion is weighted and aggregated into a composite impact score ranging from 0 to 100. All outputs are generated by the LLM, not only to identify high-potential combinations but also to reduce the risk of hallucination. Further details are provided in Appendix C.

5 Experiment Setup

5.1 Model Implementation

We employed GPT-5-2025-08-07 (OpenAI), a general-purpose LLM with strong reasoning capabilities, as the primary model in our pipeline to support structured information extraction and multi-step reasoning. The model was applied across three key stages: PICO extraction (Section 4.1), extraction and normalization of disease and drug entities (Section 4.2), and identification of candidate disease–drug combinations from emerging trends (Section 4.4).

By utilizing a consistent LLM architecture across these stages, we ensured semantic continuity and reduced the risk of formatting discrepancies that often occur when switching between different models.

5.2 Domain Selection

To provide a focused evaluation setting, this analysis is restricted to neoplastic diseases. After the entity normalization stage, only records with P disease entities mapped to the ICD-10 chapter "Neoplasms" were retained for analysis. This constraint was applied solely for experimental purposes and does not limit the general applicability of the proposed method. We are only focusing on neoplastic diseases as the representative domain due to their extensive research, well-defined disease taxonomy, and clinical importance in drug discovery. This provides a suitable testbed for evaluating the proposed approach.

5.3 Baseline

For the baseline, we use the raw biomedical abstracts with their publication years. This means the baseline approach operates on unstructured, abstract texts, requiring the LLM to implicitly extract and reason about relevant entities, relationships, and trends in the research area. As in the proposed method, the instruction to the LLM is to select the top 10 most promising P and I combinations. The generated combination should follow the normalized version of P–I. This means the output should be in ICD-10 codes for the P component and ATC codes for the I component. The generated combinations are evaluated using 4 scoring criteria: growth over the past 5 years, translational plausibility, clinical relevance, and strategic novelty. Each criterion is weighted to produce a composite impact score ranging from 0 to 100. For more details, refer to the Appendix D.

5.4 Evaluation Procedure

The evaluation is conducted using two approaches: visualization and expert interpretation. First, evaluation is performed through the interpretation of visualization, specifically heatmaps. In the heatmap, it is possible to identify where new combinations are located. This means analyzing whether the combinations lie around trending drugs or not. If

the results are located around the red region, it suggests that the model successfully captures trending drugs and may propose new diseases that can be treated using those drugs (Section 6.1). Second, a qualitative evaluation was performed by a domain expert and co-author of this paper with expertise in pharmaceutical science to determine whether the results are medically meaningful (Section 6.2).

6 Results and Discussion

6.1 Heatmap Results

We present a heatmap showing all disease–drug combinations along with the candidate combinations identified by the LLM. Out of all possible combinations of 51 diseases and 256 drugs, the proposed method identifies 32 unique disease–drug pairs. These results form the basis for the heatmap, enabling comparison between the proposed method (Figures 2a and 2c) and the baseline method (Figures 2b and 2d) in terms of identified candidate combinations. The candidate combinations are highlighted in yellow, as shown in Figure 2. We focus on candidate combinations that correspond to emerging trends (i.e., the red regions in the heatmap).

A comparison with the baseline reveals that the baseline approach identifies some candidate combinations that do not overlap with previously observed trends. Specifically, among the top 10 combinations identified by the baseline, only 2 overlap with the red regions, indicating limited alignment with emerging trends.

In contrast, the proposed method produces candidate combinations that are largely consistent with emerging trends: nearly all of the top 10 combinations overlap with the red regions. This suggests that the proposed method more effectively captures captures temporally grounded research dynamics more effectively a tabular comparison of the top 10 most plausible combinations in Table 2.

6.2 Qualitative Clinical Interpretation

Both the proposed and baseline methods show a tendency for the extracted drug–disease combinations to be concentrated in oncology, particularly in hematologic malignancies. This suggests that both methods may, to some extent, reflect the broad structure of current anticancer treatment. At the same time, however, the differences between the two methods were limited, making it difficult to draw a clear conclusion about the superiority of the proposed method based on these results alone.

When examined in more detail, the baseline included drug classes such as systemic antifungals among its top-ranked outputs, which are more commonly used for supportive care, such as infection prevention, rather than as direct anticancer therapies. This suggests that when inference is based on aggregated categories, drug classes that are not central to disease treatment may be included in the results. In contrast, the proposed method included drug classes such as corticosteroids and alkylating agents for hematologic malignancies, which are more closely aligned with the context of anticancer treatment. This may indicate that, at least in some cases, the proposed method is better at prioritizing candidates with higher clinical plausibility. However, combinations that remain difficult to interpret, such as ocular vascular disorder agents for carcinoma in situ of the digestive organs, were also observed. Therefore, it cannot be concluded that the proposed method consistently extracts meaningful candidates.

7 Conclusion & Future Work

This study aims to identify emerging areas in drug discovery by leveraging LLMs. Our basic idea is to decompose the research into `P` (patient) and `I` (intervention) combinations. Under this paradigm, given the limitation of context length, handling the entire research trend is an issue. Our solution is to represent the trend in a matrix (we all *temporal frequency delta* matrix) to incorporate it with in-context learning.

As a result, it demonstrates that while the baseline approach is more likely to suggest novel, non-overlapping 'P' and 'I' combinations, the proposed approach excels at pattern recognition within established "trending" zones. Expert interpretation confirms that, although the LLM successfully prioritizes clinically plausible treatments such as alkylating agents for hematologic malignancies, it can still produce improbable pairings, suggesting that the model currently functions best as a high-level exploratory tool for visualizing therapeutic landscapes. This expert evaluation also shows that the idea of using multiple reasoning and CIS to mitigate hallucination does not fully eliminate the risk of hallucinated outputs.

This highlights that transparency and reliability remain open challenges in LLM-based scientific discovery frameworks. Future work can also ex-

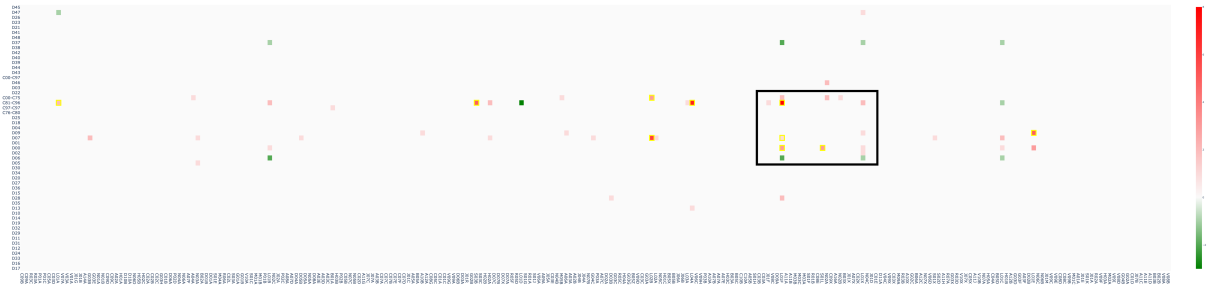
Rank	Proposed		Baseline	
	Disease (ICD-10)	Drugs (ATC)	Disease (ICD-10)	Drugs (ATC)
1	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[D07]** carcinoma in situ of other and unspecified genital organs	**[G03E]** androgens and female sex hormones in combination
2	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L04]** immunosuppressants	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L01E]** protein kinase inhibitors
3	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L02B]** hormone antagonists and related agents	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates
4	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[H02]** corticosteroids	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01E]** protein kinase inhibitors
5	**[D00]** carcinoma in situ of the oral cavity, esophagus, and stomach	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[D01]** antifungals for systemic use
6	**[D01]** carcinoma in situ of other and unspecified genital organs	**[L02B]** hormone antagonists and related agents	**[D01]**carcinoma in situ of other and unspecified digestive organs	**[L01E]** protein kinase inhibitors
7	**[D01]** carcinoma in situ of other and unspecified genital organs	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[C81–C96** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates
8	**[D07]** carcinoma in situ of other and unspecified sites	**[L01E]** protein kinase inhibitors	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[C03D]** aldosterone antagonists and other potassium-sparing agents
9	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[LO1A]** alkylating agents	**[D07]** carcinoma in situ of other and unspecified genital organs	**[G03B]** androgens
10	**[D00]** carcinoma in situ of the oral cavity, esophagus, and stomach	**[S01L]** ocular vascular disorder agents	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except for lymphoid, hematopoiesis, and related tissue	**[L04A]** immunosuppressants

Table 2: Comparison between the proposed approach and baseline approach with promising rank. Each entry consists of a code and its corresponding category name (ICD-10 for diseases and ATC for drugs). The rows with a gray background represent the combinations of the baseline and proposed approach that have the same output.

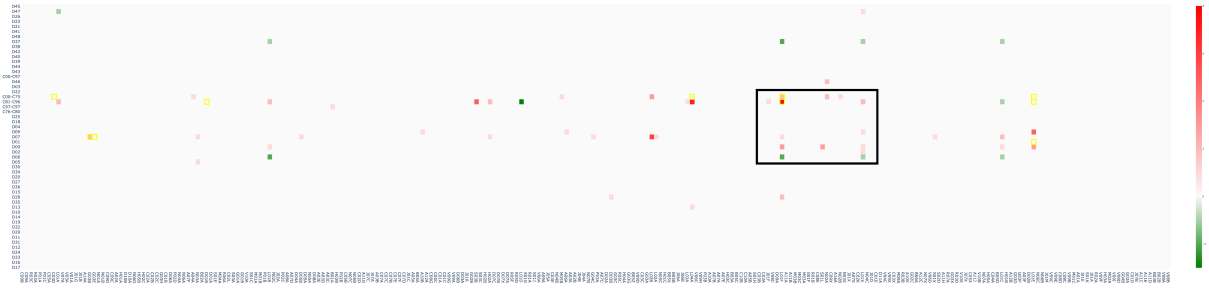
tend the analysis to other form of output such as the exact PICO output so we can evaluate the quality and the correctness of the idea generated by LLM.

8 Limitations

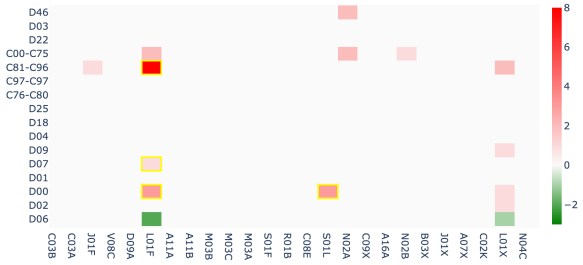
There are several limitations: First, we focus only on PubMed papers, specifically on cancer abstracts, which limits the generalizability. Second, the mapping process still needs further investigation to de-



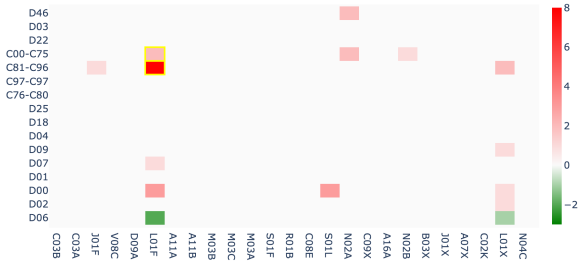
(a) Proposed method



(b) Baseline method



(c) Zoomed-in view of a selected region in (a)



(d) Zoomed-in view of a selected region in (b)

Figure 2: Heatmaps of disease (ICD-10)–drug (ATC) combinations based on temporal frequency delta (ΔF). The vertical axis represents diseases (ICD-10) and the horizontal axis represents drugs (ATC). Each cell corresponds to the ΔF value for a disease–drug pair. Red indicates combinations with increasing research activity (positive ΔF), while green indicates combinations with decreasing activity (negative ΔF); light gray indicates zero change ($\Delta F = 0$). Yellow boxes highlight top-ranked candidate combinations based on composite impact scores.

termine the optimal level of categories used for the normalization process. Third, the *temporal frequency delta* matrix ranged from -3 to 8, which may be sensitive to cumulative errors throughout the pipeline. Therefore, hallucination in each step should be considered; there may be a chance of hallucination along the way. Finally, in the evaluation framework, we only used qualitative methods that leads to missing information regarding how is the quality, the correctness, and the novelty of the new research combination. This can also be future research focusing on the evaluation using quantitative methods. Beyond these limitations, future work can be applied across more papers and sources and expanded beyond a focus on cancer. Also, the LLM models used in this research can be further explored selecting models from the medical domain is another option.

Acknowledgements

This research was supported by JST CREST JPMJCR22N1.

References

Nur Sabrina Azmi and Weng Howe Chan. 2025. Recent trends on multi-omics studies in cancer research: A bibliometric study. *International Journal of Innovative Computing*, 15(2):149–158.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6709–6738.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298.

Alireza Ghafarollahi and Markus J Buehler. 2025. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an AI co-scientist. *Preprint*, arXiv:2502.18864.

Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard A A Jonker, Sofia I R Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, Jiru Li, Dinghao Pan, Wilailack Meesawad, Richard Tzong-Han Tsai, M Janina Sarol, Gibong Hong, Airat Valiev, Elena Tutubalina, Shao-Man Lee, and 4 others. 2024. The overview of the BioRED (Biomedical Relation Extraction Dataset) track at BioCreative VIII. *Database*, 2024:baae069.

Yoonbee Kim, Yi-Sue Jung, Jong-Hoon Park, Seon-Jun Kim, and Young-Rae Cho. 2022. Drug-Disease Association Prediction Using Heterogeneous Networks for Computational Drug Repositioning. *Biomolecules*, 12(10):1497.

Zhijing Li, Yunwen Yu, Wenhao Gu, Tiantian Zhu, Haohua Song, Wenbin Guo, Xiao Yang, and Zexuan Zhu. 2025. Dual-LLM Adversarial Framework for Information Extraction from Research Literature. *Preprint*, bioRxiv:2025.09.11.675507.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. In *Proceedings of the NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Wenhao Qi, Xiaohong Zhu, Danni He, Bin Wang, Shihua Cao, Chaoqun Dong, Yunhua Li, Yanfei Chen, Bingsheng Wang, Yankai Shi, Guowei Jiang, Fang Liu, Lizzy M M Boots, Jiaqi Li, Xiajing Lou, Jiani Yao, Xiaodong Lu, and Junling Kang. 2024. Mapping knowledge landscapes and emerging trends in AI for dementia biomarkers: bibliometric and visualization analysis. *Journal of Medical Internet Research*, 26:e57830.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. 2017. MOLIERE: Automatic Biomedical Hypothesis Generation System. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1633–1642.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(509):1–14.

Ziheng Wang, Yang Zhao, and Lin Zhang. 2024. Emerging trends and hot topics in the application of multi-omics in drug discovery: A bibliometric and visualized study. *Current Pharmaceutical Analysis*, 21(1):20–32.

Norman R Williams. 2026. Old drugs, new opportunities: Advancing cancer care through repurposing. *United European Gastroenterology Journal*, 14(1):e70184.

Chao Yan, Monika E Grabowska, Alyson L Dickson, Bingshan Li, Zhexing Wen, Dan M Roden, C Michael Stein, Peter J Embí, Josh F Peterson, QiPing Feng, Bradley A. Malin, and Wei-Qi Wei. 2024. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *npj Digital Medicine*, 7(46):1–6.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Zhichao Yao, Zhiyong Lin, and Weijia Wu. 2023. Global research trends on immunotherapy in cancer: A bibliometric analysis. *Human Vaccines & Immunotherapeutics*, 19(2):2219191. PMID: 37314453.

A Details of Filtering and PICO Extraction via LLM

LLM Prompt for PICO Extraction

```
You are an expert biomedical research analyst.

The objective is to extract structured PICO/PECO
elements from biomedical abstracts while filtering out
non-human studies.

TASK:
Extract concise PICO/PECO elements and return a single
JSON object.

REQUIREMENTS:
- Output MUST contain exactly these keys: "Population",
```

```
"Intervention/Exposure", "Comparison", "Outcome", and
"Explain".
- Each field MUST contain 1–2 concise sentences using
canonical clinical phrasing.
- Do NOT include citations or bracketed identifiers.

HUMAN SUBJECT FILTER:
- If the study does NOT involve human subjects (e.g.,
animal studies, in vitro, biomedical experiments), set
ALL fields to "NA".
- Provide a clear justification in the "Explain" field.

STRICT RULES:
- Do NOT hallucinate or infer missing information.
- If any PICO element is not explicitly stated, write
"NA".
- Ensure all outputs are directly grounded in the
abstract.

WORKFLOW:
1. Perform initial screening for human subjects.
2. Identify study design, treatment arms, and
comparators.
3. Map findings into PICO/PECO components.
4. Validate that all required JSON keys are present
and correctly formatted.

OUTPUT FORMAT (JSON ONLY):

{
"Population": "...",
"Intervention/Exposure": "...",
"Comparison": "...",
"Outcome": "...",
"Explain": "..."
}

INPUT DATA:
{abstracts}
```

B Prompts for Entity Extraction & Normalization

PICO NER Prompt

```
You are a professional biomedical Named Entity
Recognition (NER) system specialized in PICO-based
extraction.

TASK:
Extract structured biomedical entities from the input
text:
- P (Population/Problem): underlying diseases or
patient conditions
- I (Intervention): therapeutic agents (drugs,
biologics)

REQUIREMENTS:
- Extract ONLY entities explicitly mentioned in the
text.
- DO NOT infer or hallucinate entities.
- Resolve abbreviations ONLY if clearly defined in the
text.
- Return ONLY the expanded (full) form.

ENTITY DEFINITIONS:
- Disease (P): pathological conditions, diagnoses,
syndromes
- Drug (I): therapeutic agents, drugs, biologics

EXCLUDE:
- Procedures, measurements, symptoms, and general terms

REASONING REQUIREMENTS:
- Each entity MUST include a "reason".
- The reason MUST be directly grounded in the text.
- Prefer exact phrase or minimal paraphrase from the
text.
- DO NOT use external knowledge.
- Maximum 15 words.

STRICT CONSTRAINTS:
- DO NOT include entities that are not directly grounded
in the text.
- DO NOT guess abbreviation expansions unless clearly
supported.
- If no entities exist, return empty lists.

VALIDATION STEP (INTERNAL):
1. Identify candidate entities directly from the text.
2. Remove duplicates and abbreviations.
3. Ensure each entity appears explicitly.
4. Ensure each reason is grounded in the text.

INPUT:
{text}

OUTPUT FORMAT (JSON ONLY):
{
"Disease": [
{"name": "...", "reason": "..."}
],
"Drug": [
{"name": "...", "reason": "..."}
]
}
```

ATC Drug Normalization Prompt

```
You are an expert biomedical terminologist performing
drug normalization using ATC classification.

TASK:
Map the input drug to the MOST SPECIFIC ATC concept
using the full hierarchy.

INPUT:
"{term}"

ONTOLOGY STRUCTURE:
- ATC hierarchy: Level 1 → Level 2 → Level 3 → Level
4 → Level 5
- Level 5 = most specific (chemical substance)

REQUIREMENTS:
- Search across ALL ATC levels
- START from Level 5 (most specific)
- If no match, move upward level-by-level
- Select the MOST SPECIFIC semantically equivalent
concept

SEMANTIC MATCHING:
- Accept synonyms, brand names, and drug combinations
- For combination drugs → choose dominant therapeutic
class

STRICT RULES:
- DO NOT choose a broader class if a specific match
exists
- DO NOT assign incorrect therapeutic class
- DO NOT hallucinate mappings
- If no valid match → return "None"

VALIDATION (INTERNAL):
1. Identify active ingredient or drug class
2. Match against candidates (all levels)
3. Start from Level 5 → move upward
4. Ensure therapeutic equivalence
5. Confirm no more specific match exists

OUTPUT FORMAT (STRICT JSON):
{
"input": "{term}",
"match": {
"code": "...",
"name": "...",
"hierarchy": [
{"level": "level1", "code": "...", "name": "..."},
{"level": "level2", "code": "...", "name": "..."},
{"level": "level3", "code": "...", "name": "..."},
{"level": "level4", "code": "...", "name": "..."},
{"level": "level5", "code": "...", "name": "..."}
]
}
}
```

```
CANDIDATES (ALL LEVELS):
{candidates}
```

ICD-10 Disease Normalization Prompt

```
You are an expert biomedical terminologist performing
disease normalization using ICD-10 (2019).

TASK:
Map the input disease to the MOST SPECIFIC matching
ICD-10 concept using the full hierarchy.

INPUT:
"{term}"

ONTOLOGY STRUCTURE:
- ICD-10 is hierarchical Level 1 → Level 2 → Level 3
→ Level 4 → Level 5 → Level 6 → Level 7
- Lower levels = more specific concepts

REQUIREMENTS:
- Search across ALL levels of ICD-10 candidates
- START from the LOWEST (most specific) level
- If no exact match at lower level, move upward
step-by-step
- Select the MOST SPECIFIC semantically equivalent
concept

SEMANTIC MATCHING:
- Accept synonyms and clinical rephrasings
- Use meaning, NOT string similarity

STRICT RULES:
- DO NOT select a broader category if a more specific
match exists
- DO NOT select a narrower category if it changes
meaning
- DO NOT hallucinate mappings
- If no valid match exists → return "None"

VALIDATION (INTERNAL):
1. Identify the core disease concept
2. Compare against candidates (all levels)
3. Start from the most specific level
4. Ensure semantic equivalence (not related, not
approximate)
5. Confirm no better (more specific) match exists

OUTPUT FORMAT (STRICT JSON):
{
"input": "{term}",
"match": {
"code": "...",
"name": "...",
"hierarchy": [
{"level": "chapter", "code": "...", "name": "..."},
{"level": "block", "code": "...", "name": "..."},
{"level": "category", "code": "...", "name": "..."},
{"level": "subcategory", "code": "...", "name": "..."}
]
}
}

CANDIDATES (ALL LEVELS):
{candidates}
```

C Prompts for Identifying Candidate Combination

Proposed Method Prompt

```
You are analyzing structured trend data derived from
publication dynamics.

Each row contains:
- Disease_ICD_L4
- Drug_ATC_L3
- Delta_Publications

DELTA = (2018–2022 publications) − (2013–2017
publications)

Meaning:
Positive → increasing research activity
Zero → no change
Negative → declining research

INPUT DATA:
{delta_json_text}

TASK:
Select the TOP 10 most promising Problem–Intervention
(P–I) combinations for future research.

SELECTION CRITERIA:
- Strong positive growth over the past 5 years
- Translational plausibility
- Clinical relevance
- Strategic novelty

SCORING SCHEME (Impact Score: 0–100):
- Growth intensity (0–30)
- Translational plausibility (0–25)
- Clinical relevance (0–25)
- Strategic novelty (0–20)

STRICT RULES:
- You MUST use EXACT labels as provided in the dataset.
- Do NOT invent, modify, or normalize labels.
- All reasoning MUST be grounded in the provided data.

OUTPUT FORMAT (JSON ONLY):

[
{
"P-Group": "...",
"I-Group": "...",
"Impact Score": 0–100,
"Insight": "Max 2 sentences reasoning"
}
]
```

D Prompts for Baseline

Baseline Prompt

```
You are analyzing biomedical abstracts along with their
publication years.

Each abstract includes a publication year. You MUST
use this temporal information to identify trends in
research activity.

The objective is to identify potential
Problem–Intervention (P–I) combinations for future
research.

TASK:
Select the TOP 10 most promising P and I combinations.

REQUIREMENTS:
- P MUST be normalized to ICD-10 Level 3 and MUST be
selected ONLY from the following valid disease labels:
{disease_labels}.
- Do NOT generate new or modified labels.
- I MUST be normalized to ATC Level 4 and MUST be
selected ONLY from the following valid drug labels:
{drug_labels}.
- Do NOT generate new or modified labels.

TEMPORAL CONSTRAINT:
- You MUST explicitly use publication years to infer
research growth.
- Compare the most recent 5 years with the earlier 5
years.
- If no temporal signal is observable, do NOT prioritize
that combination.

SCORING SCHEME (Impact Score: 0–100):
- Growth intensity (0–30)
- Translational plausibility (0–25)
- Clinical relevance (0–25)
- Strategic novelty (0–20)

STRICT RULES:
```

```
- Do NOT invent unsupported entities.
- All reasoning MUST be grounded in the abstracts AND
their publication years.

OUTPUT FORMAT (JSON ONLY):
{
"P-Group": "ICD-10 Level 3",
"I-Group": "ATC Level 4",
"Impact Score": 0-100,
"Insight":    "Max  2  sentences  including  temporal
reasoning"
}


INPUT DATA:
{abstracts}
```

E List of ICD-10 Code

ICD-10 2019 Codes

Code	Category	Code	Category
C00-C75	Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic and related tissue	C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic and related tissue
C76-C80	Malignant neoplasms of ill-defined, secondary, and unspecified sites	D48	Neoplasm of uncertain or unknown behavior of other and unspecified sites
C97-C97	Malignant neoplasms of independent (primary) multiple sites	D00	Carcinoma in situ of oral cavity, esophagus, and stomach
D01	Carcinoma in situ of other and unspecified digestive organs	D02	Carcinoma in situ of the middle ear and respiratory system
D03	Melanoma in situ	D04	Carcinoma in situ of skin
D05	Carcinoma in situ of the breast	D06	Carcinoma in situ of cervix uteri
D07	Carcinoma in situ of other and unspecified genital organs	D09	Carcinoma in situ of other and unspecified sites
D10	Benign neoplasm of the mouth and pharynx	D11	Benign neoplasm of major salivary glands
D12	Benign neoplasm of colon, rectum, anus, and anal canal	D13	Benign neoplasm of other and ill-defined parts of the digestive system
D14	Benign neoplasm of the middle ear and respiratory system	D15	Benign neoplasm of other and unspecified intrathoracic organs
D16	Benign neoplasm of bone and articular cartilage	D17	Benign lipomatous neoplasm
D18	Haemangioma and lymphangioma, any site	D19	Benign neoplasm of mesothelia tissue
D20	Benign neoplasm of soft tissue of the retroperitoneum and peritoneum	D21	Other benign neoplasms of connective and other soft tissue
D22	Melanocytic naevi	D23	Other benign neoplasms of skin
D24	Benign neoplasm of breast	D25	Leiomyoma of uterus
D26	Other benign neoplasms of uterus	D27	Benign neoplasm of ovary
D28	Benign neoplasm of other and unspecified female genital organs	D29	Benign neoplasm of male genital organs
D30	Benign neoplasm of urinary organs	D31	Benign neoplasm of eye and adnexa
D32	Benign neoplasm of meninges	D33	Benign neoplasm of brain and other parts of the central nervous system
D34	Benign neoplasm of thyroid gland	D35	Benign neoplasm of other and unspecified endocrine glands
D36	Benign neoplasm of other and unspecified sites	D37	Neoplasm of uncertain or unknown behavior of the oral cavity and digestive organs
D38	Neoplasm of uncertain or unknown behavior of the middle ear and respiratory and intrathoracic organs	D39	Neoplasm of uncertain or unknown behavior of female genital organs
D40	Neoplasm of uncertain or unknown behavior of male genital organs	D41	Neoplasm of uncertain or unknown behavior of urinary organs
D42	Neoplasm of uncertain or unknown behavior of meninges	D43	Neoplasm of uncertain or unknown behavior of brain and central nervous system
D44	Neoplasm of uncertain or unknown behavior of endocrine glands	D45	Polycythaemia vera
D46	Myelodysplastic syndromes	D47	Other neoplasms of uncertain or unknown behavior of lymphoid, hematopoietic and related tissue

F List of ATC Classification Codes

ATC Classification			
Code	**Description**	**Code**	**Description**
A01A	Stomatological Preparations	A02A	Antacids
A02B	Drugs for Peptic Ulcer and GORD	A03A	Drugs for Functional Gastrointestinal Disorders
A03B	Belladonna and Derivatives, Plain	A03C	Antispasmodics in Comb. with Psycholeptics
A03D	Antispasmodics in Comb. with Analgesics	A03E	Antispasmodics and Anticholinergics in Comb.
A03F	Propulsives	A04A	Antiemetics and Antinauseants
A05A	Bile Therapy	A05B	Liver Therapy, Lipotropics
A06A	Drugs for Constipation	A07A	Intestinal Antiinfectives
A07B	Intestinal Adsorbents	A07C	Electrolytes with Carbohydrates
A07D	Antipropulsives	A07E	Intestinal Antiinflammatory Agents
A07F	Antidiarrheal Microorganisms	A07X	Other Antidiarrheals
A08A	Antiobesity Preparations	A09A	Digestives, Incl. Enzymes
A10A	Insulins and Analogues	A10B	Blood Glucose Lowering Drugs, Excl. Insulins
A10X	Other Drugs Used in Diabetes	A11A	Multivitamins, Combinations
A11B	Multivitamins, Plain	A11C	Vitamin A and D, Incl. Combinations
A11D	Vitamin B1, Plain and Comb.	A11E	Vitamin B-Complex, Incl. Combinations
A11G	Ascorbic Acid (Vitamin C), Incl. Comb.	A11H	Other Plain Vitamin Preparations
A11J	Other Vitamin Products, Combinations	A12A	Calcium
A12B	Potassium	A12C	Other Mineral Supplements
A14A	Anabolic Steroids	A16A	Other Alimentary Tract and Metabolism Products
B01A	Antithrombotic Agents	B02A	Antifibrinolytics
B02B	Vitamin K and Other Hemostatics	B03A	Iron Preparations
B03B	Vitamin B12 and Folic Acid	B03X	Other Antianemic Preparations
B05A	Blood and Related Products	B05B	I.V. Solutions
B05C	Irrigating Solutions	B05D	Peritoneal Dialytics
B05X	I.V. Solution Additives	B05Z	Hemodialytics and Hemofiltrates
B06A	Other Hematological Agents	C01A	Cardiac Glycosides
C01B	Antiarrhythmics, Class I and III	C01C	Cardiac Stimulants Excl. Cardiac Glycosides
C01D	Vasodilators Used in Cardiac Diseases	C01E	Other Cardiac Preparations
C02A	Antiadrenergic Agents, Centrally Acting	C02B	Antiadrenergic Agents, Ganglion-Blocking
C02C	Antiadrenergic Agents, Peripherally Acting	C02D	Arteriolar Smooth Muscle, Agents Acting On
C02K	Other Antihypertensives	C02L	Antihypertensives and Diuretics in Comb.
C03A	Low-Ceiling Diuretics, Thiazides	C03B	Low-Ceiling Diuretics, Excl. Thiazides
C03C	High-Ceiling Diuretics	C03D	Aldosterone Antagonists and K-Sparing Agents
C03E	Diuretics and K-Sparing Agents in Comb.	C03X	Other Diuretics
C04A	Peripheral Vasodilators	C05A	Agents for Hemorrhoids (Topical)
C05B	Antivaricose Therapy	C05C	Capillary Stabilizing Agents
C05X	Other Vasoprotectives	C07A	Beta Blocking Agents
C07B	Beta Blocking Agents and Thiazides	C07C	Beta Blocking Agents and Other Diuretics
C07D	Beta Blocking Agents, Thiazides and Diuretics	C07E	Beta Blocking Agents and Vasodilators
C07F	Beta Blocking Agents, Other Comb.	C08C	Selective Calcium Channel Blockers (Vascular)
C08D	Selective Calcium Channel Blockers (Cardiac)	C08E	Non-Selective Calcium Channel Blockers
C08G	Calcium Channel Blockers and Diuretics	C09A	ACE Inhibitors, Plain
C09B	ACE Inhibitors, Combinations	C09C	Angiotensin II Receptor Blockers (ARBs)
C09D	ARBs, Combinations	C09X	Other Renin-Angiotensin System Agents
C10A	Lipid Modifying Agents, Plain	C10B	Lipid Modifying Agents, Combinations
D01A	Antifungals for Topical Use	D01B	Antifungals for Systemic Use
D02A	Emollients and Protectives	D02B	Protectives Against UV-Radiation
D03A	Cicatrizants	D03B	Enzymes
D04A	Antipruritics	D05A	Antipsoriatics for Topical Use
D05B	Antipsoriatics for Systemic Use	D06A	Antibiotics for Topical Use
D06B	Chemotherapeutics for Topical Use	D07A	Corticosteroids, Plain
D07B	Corticosteroids, Comb. with Antiseptics	D07C	Corticosteroids, Comb. with Antibiotics
D07X	Corticosteroids, Other Combinations	D08A	Antiseptics and Disinfectants
D09A	Medicated Dressings	D10A	Anti-Acne Preparations for Topical Use
D10B	Anti-Acne Preparations for Systemic Use	D11A	Other Dermatological Preparations
G01A	Antiinfectives and Antiseptics (Vaginal)	G01B	Antiinfectives with Corticosteroids
G02A	Uterotonics	G02B	Contraceptives for Topical Use
G02C	Other Gynecologicals	G03A	Hormonal Contraceptives for Systemic Use
G03B	Androgens	G03C	Estrogens
G03D	Progestogens	G03E	Androgens and Female Sex Hormones in Comb.
G03F	Progestogens and Estrogens in Combination	G03G	Gonadotropins and Ovulation Stimulants
G03H	Antiandrogens	G03X	Other Sex Hormones
G04B	Urologicals	G04C	Drugs Used in Benign Prostatic Hypertrophy
H01A	Anterior Pituitary Lobe Hormones	H01B	Posterior Pituitary Lobe Hormones
H01C	Hypothalamic Hormones	H02A	Corticosteroids for Systemic Use, Plain
H02B	Corticosteroids for Systemic Use, Comb.	H02C	Antiadrenal Preparations
H03A	Thyroid Preparations	H03B	Antithyroid Preparations
H03C	Iodine Therapy	H04A	Glycogenolytic Hormones
H05A	Parathyroid Hormones	H05B	Anti-Parathyroid Agents
J01A	Tetracyclines	J01B	Amphenicols
J01C	Beta-Lactam Antibacterials, Penicillins	J01D	Other Beta-Lactam Antibacterials

ATC Classification

Code	Description	Code	Description
J01E	Sulfonamides and Trimethoprim	J01F	Macrolides and Lincosamides
J01G	Aminoglycoside Antibacterials	J01M	Quinolone Antibacterials
J01R	Combinations of Antibacterials	J01X	Other Antibacterials
J02A	Antimycotics for Systemic Use	J04A	Drugs for Treatment of Tuberculosis
J04B	Drugs for Treatment of Lepra	J05A	Direct Acting Antivirals
J06A	Immune Sera	J06B	Immunoglobulins
J07A	Bacterial Vaccines	J07B	Viral Vaccines
J07C	Bacterial and Viral Vaccines, Combined	J07X	Other Vaccines
L01A	Alkylating Agents	L01B	Antimetabolites
L01C	Plant Alkaloids	L01D	Cytotoxic Antibiotics
L01E	Protein Kinase Inhibitors	L01F	Monoclonal Antibodies and ADCs
L01X	Other Antineoplastic Agents	L02A	Hormones and Related Agents
L02B	Hormone Antagonists	L03A	Immunostimulants
L04A	Immunosuppressants	M01A	Antiinflammatory (NSAIDs)
M01B	Antiinflammatory Agents in Comb.	M01C	Specific Antirheumatic Agents
M02A	Topical Products for Joint/Muscular Pain	M03A	Muscle Relaxants, Peripherally Acting
M03B	Muscle Relaxants, Centrally Acting	M03C	Muscle Relaxants, Directly Acting
M04A	Antigout Preparations	M05B	Drugs Affecting Bone Structure
M09A	Other Musculo-Skeletal Drugs	N01A	Anesthetics, General
N01B	Anesthetics, Local	N02A	Opioids
N02B	Other Analgesics and Antipyretics	N02C	Antimigraine Preparations
N03A	Antiepileptics	N04A	Anticholinergic Agents
N04B	Dopaminergic Agents	N04C	Other Antiparkinson Drugs
N05A	Antipsychotics	N05B	Anxiolytics
N05C	Hypnotics and Sedatives	N06A	Antidepressants
N06B	Psychostimulants and Nootropics	N06C	Psycholeptics and Psychoanaleptics Comb.
N06D	Anti-Dementia Drugs	N07A	Parasympathomimetics
N07B	Drugs Used in Addictive Disorders	N07C	Antivertigo Preparations
N07X	Other Nervous System Drugs		
P01A	Agents Against Amoebiasis	P01B	Antimalarials
P01C	Agents Against Leishmaniasis	P02B	Antitrematodals
P02C	Antinematodal Agents	P02D	Anticestodals
P03A	Ectoparasiticides	P03B	Insecticides and Repellents
R01A	Nasal Decongestants (Topical)	R01B	Nasal Decongestants (Systemic)
R02A	Throat Preparations	R03A	Adrenergics, Inhalants
R03B	Other Inhalants for Obstructive Airway	R03C	Adrenergics for Systemic Use
R03D	Other Systemic Drugs for Obstructive Airway	R05C	Expectorants
R05D	Cough Suppressants	R05F	Cough Suppressants and Expectorants
R06A	Antihistamines for Systemic Use	R07A	Other Respiratory System Products
S01A	Antiinfectives (Ophthalmic)	S01B	Antiinflammatory (Ophthalmic)
S01C	Antiinflam. and Antiinfectives Comb.	S01E	Antiglaucoma and Miotics
S01F	Mydriatics and Cycloplegics	S01G	Decongestants and Antiallergics
S01H	Local Anesthetics (Ophthalmic)	S01J	Diagnostic Agents (Ophthalmic)
S01K	Surgical Aids (Ophthalmic)	S01L	Ocular Vascular Disorder Agents
S01X	Other Ophthalmologicals	S02A	Antiinfectives (Otic)
S02B	Corticosteroids (Otic)	S02C	Corticosteroids and Antiinfectives Comb.
S02D	Other Otologicals	S03A	Antiinfectives (Ophth/Otic)
S03B	Corticosteroids (Ophth/Otic)	S03C	Corticosteroids/Antiinfectives Comb.
V01A	Allergens	V03A	All Other Therapeutic Products
V04C	Other Diagnostic Agents	V06A	Diet Formulations for Obesity
V06C	Infant Formulas	V06D	Other Nutrients
V07A	All Other Non-Therapeutic Products	V08A	X-Ray Contrast Media, Iodinated
V08B	X-Ray Contrast Media, Non-Iodinated	V08C	MRI Contrast Media
V08D	Ultrasound Contrast Media	V09A	CNS (Radiopharmaceuticals)
V09B	Skeleton (Radiopharmaceuticals)	V09C	Renal (Radiopharmaceuticals)
V09D	Hepatic (Radiopharmaceuticals)	V09E	Respiratory (Radiopharmaceuticals)
V09F	Thyroid (Radiopharmaceuticals)	V09G	Cardiovascular (Radiopharmaceuticals)
V09H	Inflammation Detection (Radiopharma)	V09I	Tumour Detection (Radiopharmaceuticals)
V09X	Other Diagnostic Radiopharma	V10A	Antiinflammatory (Therapeutic Radiopharma)
V10B	Pain Palliation – Bone Seeking	V10X	Other Therapeutic Radiopharmaceuticals

G Example of Extracted PICO Components

Table 3: An example of PICO components extracted from abstract, drugs and disease extracted from `P` & `I` components, and results mapped to ICD-10 and ATC

Field	Content
PMID	31171494
Year	2021
Abstract	Older patients are one of the most relevant sub-groups of patients with breast cancer and will only gain in importance as demographic transition unfolds. Their management, in both the early and advanced settings, should take into consideration specific clinical needs and is made more difficult by the limited availability of evidence on the efficacy and safety of standard treatment regimens in older patients. At the root of this situation is the low rate of participation of older patients in clinical trials, often due to age limits for inclusion, and limitations on the participation of persons with significant comorbidities or organ dysfunction. Although this has begun to change in recent years, most agents currently in use have not been tested in a substantial number of older patients. This includes targeted agents that have changed prognosis in breast cancer.
Population	Older adults with early or advanced breast cancer, a subgroup underrepresented in clinical trials.
Intervention	Use of targeted agents including HER2-targeted therapies (trastuzumab, lapatinib, pertuzumab, T-DM1, neratinib), CDK4/6 inhibitors (palbociclib, ribociclib, abemaciclib), bevacizumab, everolimus, and PARP inhibitors (olaparib, talazoparib).
Comparison	NA
Outcome	Efficacy and safety profiles with an older-specific toxicity and tolerability context to inform treatment decisions.
Explain	Narrative review summarizing efficacy and safety data; no explicit comparator specified.
P_Disease	Breast cancer
I_Drugs	trastuzumab, lapatinib, pertuzumab, trastuzumab emtansine, neratinib, CDK4/6 inhibitors, palbociclib, ribociclib, abemaciclib, bevacizumab, everolimus, PARP inhibitors, olaparib, talazoparib
P_Code	C00–C75
P_Name	Malignant neoplasms of specified sites (excluding lymphoid and related tissue)
I_Code	L01F, L01E, L01X, L04A
I_Name	Monoclonal antibodies, Protein kinase inhibitors, Other antineoplastic agents, Immunosuppressants

H Entity Normalization by Ontology Mapping

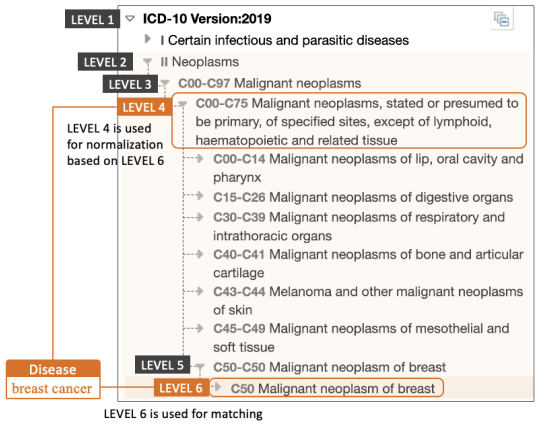


Figure 3: ICD-10 code mapping

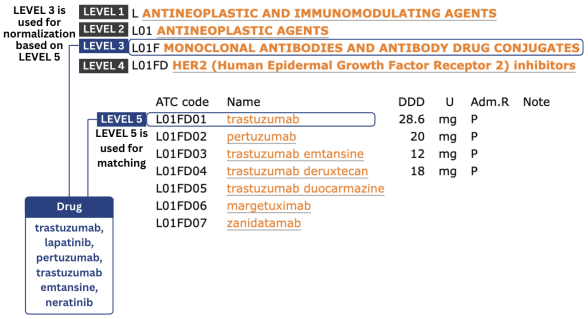


Figure 4: ATC code mapping

FHexchange: Resources for Family Health History Extraction and Normalization From Consumer Dialog Sources

Michelle H. Nguyen[1*], Nidhi Soley[1], Ayah Zirikly[2,3], João Sedoc [4], Casey Overby Taylor [1]
[1] Institute for Computational Medicine, Johns Hopkins University
[2] Center for Language and Speech Processing, Johns Hopkins University
[3] Department of Computer Science, George Washington University
[4] Stern School of Business, New York University
[*]**Correspondence**: {mnguye79} @jhmi.edu

Abstract

Family health history (FHx) offers insight into a person's health and disease risk, but it is largely held within free-text clinical formats that require processing for maximal utility of the data. The rapid deployment of ambient AI scribes and conversational agents in clinical settings necessitates evaluation on dynamic patient-clinician and patient-agent dialogs. To address this gap, we introduce two new datasets of patient FHx dialog documents designed to benchmark information extraction and entity linking. Distinct from clinician-entered datasets, patient-reported dialog data has its own semantic and content characteristics, which need to be studied for more patient-centered healthcare. We contribute a publicly available resource called *FHexchange*, with new annotations for family members, clinical observations, related entities, and standardized UMLS CUIs, offering the clinical NLP community a robust evaluation bed for emerging generative AI tools.

1 Introduction

Family health history (FHx) is an important health data type that gives insight into genetic, environmental, and behavioral factors that contribute to a person's disease risk. This information may be collected in a variety of ways, through informational interviews with genetic counselors, primary care providers, and other clinicians, or through asynchronous surveys and chatbots (Wildin et al., 2021). The data collection methods also determine several features of the FHx data, including document and data structure as well as the granularity of data. When FHx information has low granularity (i.e., unstructured symptom descriptions missing standardized concept mappings), it can limit how useful the data are for downstream prediction tasks and clinical decision support (Moon et al., 2022). Thus, to overcome the problem of low granularity, we can extract FHx information in a standardized, structured format and normalize them to standard vocabularies when possible (Chen et al., 2015).

Much of the collected FHx data still remains within free-text sources; thus, natural language processing (NLP) approaches are necessary to extract this important data (Zhou et al., 2014). Previous work has explored FHx information extraction tasks and relation linking tasks in both synthetic and real-world electronic health record data sources (Shen et al., 2021; Bill et al., 2014; Rybinski et al., 2021; Silva et al., 2020; Goryachev et al., 2008; Zhan et al., 2021), as well as from FHx questionnaires (Shing et al., 2021) using a variety of methods. Methods have included neural language model embeddings with deep learning architectures, rule-based systems, as well as hybrid approaches. These approaches were largely developed on a few public datasets, but more often on proprietary datasets that are not available to the larger clinical NLP research community. Additionally, there are few FHx datasets that provide normalized concepts to support the study of entity linking to clinical codes.

As new AI tools are deployed in the clinic, datasets that reflect new modalities, such as patient-clinician dialogs for ambient AI scribes or conversational AI interactions, are necessary. To support the assessment of FHx extraction for emerging data types, we are contributing the following:

- a newly annotated and normalized resource called *FHexchange* for FHx entity extraction and relation extraction that is publicly available,
- with additional National Library of Medicine's Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) (U.S. National Library of Medicine) for entity linking benchmarking (that will be accessible for eligible researchers with a UMLS license)

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1014–1028
July 3-4, 2026 ©2026 Association for Computational Linguistics

based on two existing dialog corpora: 104 FHx chatbot-user dialogs from (Nguyen et al., 2024), referred to as *FHexchange-KIT*, and 154 FHx-related dialogs from the openly available MTS-Dialog dataset (Ben Abacha et al., 2023), referred to as *FHexchange-MTS*.

Enabled by *FHexchange*, we discuss the following research topics:

- The effect of collection methods on the characteristics and condition coverage of FHx data
- The potential for collection methods to inform what FHx data patients choose to disclose
- The utility of these datasets to support the development of new clinical AI tools.

2 Related Work

2.1 Existing FHx guidelines and schema

In 2008, the Family Health History Multi-Stakeholder Workgroup of the American Health Information Community convened to formalize a core set of FHx elements. The final core set included person-level attributes (identification, ages, ages/dates at death, cause of death, ethnicity/race, sex, multiple-birth status, adoptive status, relevant genetic/genomic test results), condition-level attributes (disorders list, notation for unknown and/or uncertain histories, and relevant age ranges), and other attributes (consanguinity, identification of biological parents, sensitive data, and integration with other electronic health record data elements) (Feero et al., 2008). This work contributes additional high-quality annotated data for FHx data extraction and concept normalization.

As a part of the Social and Family History - Extraction, Representation, and Evaluation (SFHERE) Project, Chen et al. conducted a multi-source investigation of FHx collection tools and data elements to compile a comprehensive set of relevant FHx entities. They found a wide range of FHx elements represented across three distinct clinical note data sources (MTSamples, University of Pittsburgh Medical Center, and Fletcher Allen Health Care), patient-facing data collection tools (survey of eight tools), and FHx specifications (survey of seven specifications). Through iterative revisions, they designed a highly detailed annotation guideline to direct the manual labeling of clinical notes. The final version of entities and attributes included 13 unique subject categories, nine unique observation categories, four age/date-related categories, and two categories for other FHx-related information. The annotation resources for the SFHERE Project are publicly available (SFHERE Team, 2017).

Rama et al. (2018) also explored an iterative annotation process for synthetic Norwegian FHx clinical notes. The authors annotated for family members, self (which represents the proband), the condition entries (including diseases, diagnoses, mutations, test results, treatments, and living statuses), and medical events (including acute occurrences). Modifying entities captured included side of the family, age, Negateds, amounts (e.g., *three* sisters, *two* strokes), and temporal expressions (e.g., had a heart attack *two years ago*), which were then related to previously described clinical entities (Rama et al., 2018).

2.2 Existing datasets

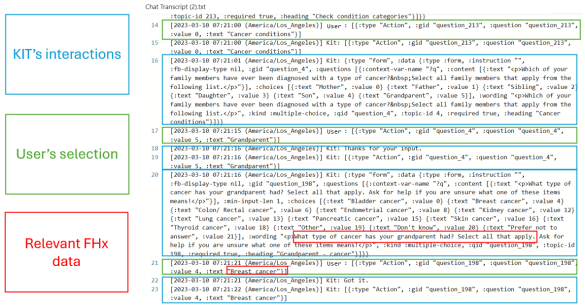
A summary of datasets that were specifically annotated for FHx mentions can be found in Table 1. Historically, FHx extraction research has heavily relied on clinician-authored or de-identified notes. The MTSamples dataset was used by several groups to explore FHx extraction (Lewis et al., 2011; Bill et al., 2014; Chen et al., 2015); however, only 100 documents are publicly accessible as of 2026 from the SFHERE Project (2014). The 2019 n2c2/OHNLP and BioCreative/OHNLP challenge datasets (Shen et al., 2021; Rastegar-Mojarad et al., 2018) also provided foundational benchmarks for extracting entities and relations, from highly detailed genetics-specialty-style notes. Wang et al. reannotated the datasets to correct errors in the original version as well as to enhance the data by normalizing observations to SNOMED-CT codes and creating a FHx lexical resource (Wang et al., 2023). Both corpora are only accessible through a data use agreement (DUA). Rama et al. continued their work and released the dataset of 477 cardiac-related FHx sentences (in Norwegian) in 2021 (Brekke et al., 2021), which is freely accessible.

Recent work has explored consumer-reported and patient-involved datasets. Azab et al. (2019) created a FHx dataset of 228 patient-facing questionnaires from genetic counseling sessions and crowd-sourced participants. This dataset was very rich and information-dense, containing over 20,000 FHx relations; however, the dataset is no longer accessible. Saley et al. (2024) released MediTOD in 2024 based off a publicly available set of patient-

Dataset	Type/Source	Language	FHx annotation types	Disease focus	Size of FHx-related content	FHx annotation access*
Clinician-authored sources						
MTSamples (MTSamples)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (such as side of family, Negated), and relation extraction.	Variety	491-3940 documents	Freely accessible for 100 documents (SFHERE Project, 2014)
2019 n2c2/OHNLP FH (Shen et al., 2021); BioCreative/OHNLP 2018 (Rastegar-Mojarad et al., 2018)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (side of family, Negated), and relation extraction.	Variety/ Genetics	216 documents	DUA required (99 documents, training set)
Brekke et al. (2021)	Synthetic clinical note	Norwegian	Clinical entity identification (family members, event, conditions), entity properties (ages, side of family, Negated, amount, temporal), and relation extraction	Cardiology	477 sentences	Freely accessible
2019 n2c2/OHNLP FH; BioCreative/OHNLP 2018 **re-annotated** (Wang et al., 2023)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (side of family, Negated), relation extraction, normalized concepts (SNOMED CT).	Variety/ Genetics	216 documents	DUA required
Consumer-reported sources						
Azab et al. (2019)	Real patient-facing questionnaire	English	Clinical entity identification (family members, illnesses, cancer, cancer type), entity properties (living status, sex, ages), and relation extraction.	Cancer	228 questionnaires	No longer accessible
MediTOD (Fareez et al., 2022; Saley et al., 2024)	Synthetic patient-clinician dialog	English	Clinical entity identification (family members, observations), entity properties (status, Negated), relation extraction, normalized concepts (UMLS CUI).	Respiratory	214 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)
FHexchange-MTS, from MTS-DIALOG subset (Ben Abacha et al., 2023)	Synthetic patient-clinician dialog	English	Clinical entity identification (family members, observations, living status, ages), entity properties (e.g., side of family, Negated), relation extraction, normalized concepts (UMLS CUI)	Variety	154 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)
FHexchange-KIT, from KIT-user dialogs (Nguyen et al., 2024)	Real user-chatbot interaction transcript	English	Clinical entity identification (family members, observations), relation extraction, normalized concepts (UMLS CUI)	Variety	104 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)

Table 1: Snapshot of released datasets by source type with FHx entity, relation annotations, and normalized concepts, with our contributions highlighted , *access as of 2026

clinician dialogs from Fareez et al. (2022). The authors labeled dialogs for conversational attributes, such as the intent of the dialog turn, as well as topic areas, termed by the authors as *slots*, which could be information regarding symptoms, collection of patient medical history, or collection of family medical history, which was of particular interest for the present work. There are 214 dialogs containing FHx mentions, and the authors primarily labeled for family member, condition, status and Negated, leaving other FHx details in a semi-structured text field. To support additional evaluation of clinical



(a) KIT Transcript from a user interaction. In the red box, we can identify that the user has indicated that their family has a history of cancer conditions, specifically that their grandparent had breast cancer.

Doctor: Tell me about your family. Did anyone have any medical issues?
Patient: My father is alive. He is sixty nine right now. He did have some strokes in the past.
My mom on the other hand died at the age of sixty two of many kinds of heart diseases.
Doctor: What about your grandparents?
Patient: My maternal grandfather he also had heart issues and he died of a heart attack.
Both he and my maternal grandma had some heart diseases. Their heart use to beat pretty fast.

FamilyMember LivingStatus SideoftheFamily Observation AgeatDeath Age

(b) MTS-Dialog sample patient-doctor conversation. The annotated entities are highlighted.

Figure 1: Examples of KIT and MTS-Dialog data

concepts, Saley et al. (2024) mapped observations to UMLS codes using QuickUMLS (Soldaini and Goharian).

We contribute *FHexchange*, a new dataset consisting of two additional consumer-reported sources derived from: (1) the MTS-Dialog dataset (Ben Abacha et al., 2023) and (2) KIT-user dialogs (Nguyen et al., 2024). The full MTS-Dialog dataset consists of 1701 synthetic patient-doctor conversation documents and summarized clinical notes for these dialogs. Because the original dataset was designed to support clincal note summarization, the authors did not annotate for specific elements such as FHx. To construct *FHexchange-MTS*, we selected a set of 154 MTS-Dialog dialogs that contained FHx mentions for further FHx annotation, a sample dialog can be seen in Figure 1b. For the second subdataset, *FHexchange-KIT*, we annotated 104 KIT-user dialogs from a crowdsourced study to assess chatbot-based data collection from a chatbot named KIT (Nguyen et al., 2024). The questionnaire was adapted from the NIH *AllofUs* family history survey and covered intergenerational FHx for 11 pre-defined condition categories (The All of Us Research Program Investigators, 2019). Figure 1a shows an example of an *FHexchange-KIT* excerpt.

3 *FHexchange* dataset

3.1 Annotation

The annotation guideline was created by referencing existing schema (Chen et al., 2015) for a subset of terms and further development with clinical experts (genetic counselors). The annotation guideline was iteratively refined by annotating a sample set of clinical notes with guidance from two genetic counselors, each with over ten years of clinical experience.

The *FHexchange-MTS* data was annotated for the following concepts: *FamilyMember, Observation, SideoftheFamily, LivingStatus, Age, AgeofOnset, AgeofDeath, CauseofDeath, and Negated*, following a schema further described in Appendix A. The *FHexchange-KIT* data consisted of user-entered health conditions for a specified set of *FamilyMembers*: Father, Mother, Sibling, Grandparent, Son, and Daughter. Because of this, the annotations were limited to *FamilyMember* and *Observation*. Two Masters-level individuals (both native English speakers) trained in Biomedical Engineering completed the clinical entity annotations. Both annotators had overlap on the full corpus.

UMLS concept normalization was first conducted using the python library, scispacy's UMLS entity linker, with a link score (calculated as a nearest neighbor search percentage between the candidate text and the UMLS CUIs) (Neumann et al., 2019). Only matches with a link score of over 0.9 were retained. One human annotator also normalized all concepts for both *FHexchange-MTS* and *FHexchange-KIT* clinical observations for comparison with automated code assignments.

3.2 *FHexchange* dataset summary

We report interrater agreement using Cohen's kappa for the presence of *Observations* and *FamilyMembers*. We also report percent agreement for both datasets. For the *FHexchange-KIT* subdataset, we report percent agreement on a document-level as the annotations were limited to a fixed set of *FamilyMembers*. Percent agreement was calculated as the proportion of overlapping observations between annotators across six predefined *Family-*

Members, while full agreement strictly required perfectly identical observation sets for every family member within the transcript.

To report coverage of FHx conditions within the dataset, we performed further linking of UMLS concepts to Semantic Types and ICD-10 Body Systems using UMLS API calls with UMLS version 2025AB. We also reported the distribution of UMLS semantic types and ICD-10 body systems present. All computation was performed using python version 3.11.

3.3 Example baselines with zero-shot and few-shot large language model structured output approaches

We demonstrate the value of the dataset by providing an example baseline for structured FHx information extraction, relation extraction, and entity linking using the following pre-trained language models with structured outputs: Microsoft Foundry Azure OpenAI zero-shot GPT-4o (Version 2024-11-20) (Microsoft, a), GPT-5-mini (Version 2025-08-07) (Microsoft, b), and Microsoft Azure Llama-3.3-70B-Instruct (Version 9) (Microsoft, c). We aimed to select proprietary models (flagship model versus lightweight reasoning model) and an open-weight model. Although we did not include a clinical model baseline, we included an open-weight model to represent a solution that could be locally deployed, considering data privacy and compliance with healthcare regulations. We created a structured Pydantic (Colvin et al., 2026) schema to prompt the model to reason through FHx extraction (with a *reasoning* field as the first entry in the schema) of the relevant FHx annotated entities where available (*FamilyMember*, *Observation*, *AgeofOnset*, *SideoftheFamily*, *LivingStatus*, *Age*, *AgeofDeath*, *CauseofDeath*, CUI, and *Negated*). Model, prompt details, and schema description can be found in the Appendix, and additional details are available on the dataset Github repository.

To handle the structural differences between *FHexchange-KIT* and *FHexchange-MTS*, the schema utilized optional typing, allowing the models to output `null` for variables unsupported by the text without breaking the JSON generation. Additionally, prompt constraints were applied to standardize clinical concepts, such as negated histories (e.g., forcing the extraction of the base condition with a *Negated: true* boolean flag) and restricting categorical modifiers to predefined values (such as Maternal, Paternal, or Unknown, for *SideoftheFamily*).

After normalizing the model output, evaluation was performed by reporting precision, recall, and F1 for all entities extracted and all relations extracted. Matches for *Observations* and *FamilyMembers* were evaluated strictly based off perfect token-level matches compared to the gold standard annotations.

Metric	Value
FHexchange-MTS	
Binary Presence (κ)	0.827
Strict match (κ)	0.300
Lenient match (κ)	0.519
FHexchange-KIT	
Binary Presence (κ)	0.990
Perfect match (%)	63.5%
Partial match (%)	36.5%

Table 2: Inter-annotator agreement across datasets. For *FHexchange-MTS*, we report phrase-level κ representing strict and lenient text span matches. For *FHexchange-KIT*, we report κ for the presence of *FamilyMember* and *Observation* fields and raw document-level percent agreement for exact sets of conditions. Additional agreement details for *Observation-Family Member* relations can be seen in Appendix Table 9.

Concept	*FHexchange-MTS*	*FHexchange-KIT*
Dataset Statistics		
Documents	154	104
Dialog turns	856	5156
Annotation Counts		
FamilyMember	265	249
Observation	488	720
SideoftheFamily	24	–
LivingStatus	283	–
Age	20	–
AgeofOnset	12	–
AgeofDeath	59	–
CauseofDeath	70	–
CUI	424	679
Negated	112	–
Relations (Total)	1280	1399

Table 3: Annotation statistics for *FHexchange-MTS* and *FHexchange-KIT*

4 Results

Appendix Table 2 details the results of annotator agreement for the two datasets. For *FHexchange-MTS*, Cohen's kappa for the overlap regarding the presence of the entities: *FamilyMember* type, *LivingStatus* type, or *SideoftheFamily* type and *Observation*, was high (0.827, indicating a near perfect

Rk	Condition	N	Sample Expressions
FHexchange-MTS			
1	Diabetes mellitus	46	*diabetes, has sugar*
2	Heart diseases	43	*heart disease, heart issues, heart problems*
3	Hypertension	26	*high blood pressure, hypertension*
4	Cerebrovascular accident	23	*stroke, CVA*
5	Healthy	20	*healthy, no conditions*
FHexchange-KIT			
1	High blood pressure - hypertension	53	-
2	Depression	37	-
3	Heart attack	28	-
4	High cholesterol	26	-
5	Anxiety reaction/panic disorder	21	-

Table 4: Top 5 conditions by subdataset with sample patient expressions. Top conditions for *FHexchange-KIT* did not contain any variance, so no sample expressions are listed.

match) (Landis and Koch, 1977). When comparing a strict phrase-level match, where tokens must perfectly overlap, the kappa value was low (0.300); however, when allowing for fuzzy matches for a lenient phrase-level match, the agreement improves between annotators (0.519). Appendix B provides examples of errors between annotators. The performance in the results for the strict match were lower due to differences in the boundaries annotators notated (Table 9 provides additional details regarding categories of interannotator disagreement). For example, for age at death, one annotator specified *AgeofDeath* was "40s", but another specified this was "died in 40s." For this example, we removed all modifying text related to *AgeofDeath*, to only retain "40s." For *FHexchange-KIT*, there was high agreement for the presence of *FamilyMember* and *Observation* with a Cohen's κ of 0.990. For document-level agreement, 63.5% of the documents were perfect matches between annotators.

Table 3 contains the annotation statistics across both datasets. *FHexchange-MTS* contains 856 dialog turns, and *FHexchange-KIT* contains 5156 dialog turns.

Table 4 presents the top conditions and frequencies after mapping *Observations* to CUIs across both datasets. Sample expressions are provided for *FHexchange-MTS* to reflect the variations in patient descriptors for diseases. The two subdatasets both have hypertension in the top five conditions.

When observing higher-order categories of conditions, we identified eight unique Semantic Types of mapped CUIs from the *FHexchange-KIT* subdataset and 13 Semantic Types for *FHexchange-MTS*, as seen in Table 5. We found 12 unique ICD-10 Body systems for *FHexchange-KIT* and 11

Category	FHex-KIT	FHex-MTS
Unique CUIs	**114**	**121**
Semantic Types (FHex-KIT: 8, FHex-MTS : 13)		
Disease or Syndrome	68	54
Neoplastic Process	22	30
Mental/Behavioral Dysfunction	11	10
Pathologic Function	6	6
Finding	3	7
Sign or Symptom	2	4
Anatomical Abnormality	1	3
Congenital Abnormality	1	–
Therapeutic/Preventive Procedure	–	3
Qualitative Concept	–	1
Unknown	–	1
Injury or Poisoning	–	1
Event	–	1
Medical Device	–	1
ICD-10 Body Systems (FHex-KIT: 12, FHex-MTS-Dialog: 11)		
No Direct ICD-10 Mapping	26	48
Neoplasms (Oncology)	16	18
Other/General	12	7
Nervous System	12	12
Digestive System	11	3
Mental and Behavioral	9	6
Endocrine/Nutritional/Metabolic	8	6
Circulatory System	7	12
Musculoskeletal System	7	4
Respiratory System	3	4
Genitourinary System	2	2
Blood/Blood-forming Organs	1	-

Table 5: UMLS semantic types and ICD-10 body system distributions across *FHexchange-KIT (FHex-KIT)* and *FHexchange-MTS (FHex-MTS)*

Body systems for *FHexchange-MTS*.

In Table 6, we report baseline performance of the two proprietary pre-trained language models, Azure OpenAI GPT-4o and GPT-5-mini with zero-shot and few-shot in-context learning approaches, to extract family history information and to perform relation extraction and UMLS entity linking. The full experimental results with performance from all three models (including Azure Llama 3.3 70B Instruct) can be seen in Appendix Section D. Zero-shot prompting performance on extracting *FamilyMembers* for both subdatasets is high for both models at baseline (GPT-4o: F1 = 0.994 for *FHexchange-KIT*, F1 = 0.851 for *FHexchange-MTS*; GPT-5-mini: F1 = 0.972 for *FHexchange-KIT*, F1 = 0.774 for *FHexchange-MTS*). However, performance declined for both with few-shot prompting for the *FHexchange-KIT*

		FM	Obs	Rel	CUI	Neg	Living	Side
FHexchange-KIT								
GPT-4o	ZS	.994 (.99/1.0)	.578 (.53/.64)	.632 (.66/.61)	.354 (.36/.35)	.573 (.52/.64)	—	—
	FS	.986 (.98/.99)	.638 (.58/.71)	.702 (.71/.69)	.373 (.37/.38)	.039 (.08/.03)	—	—
GPT-5-mini	ZS	.972 (.97/.97)	.726 (.66/.81)	.824 (.85/.80)	.115 (.48/.07)	.714 (.65/.79)	—	—
	FS	.891 (.97/.82)	.649 (.63/.67)	.711 (.79/.65)	.255 (.44/.18)	.000	—	—
FHexchange-MTS								
GPT-4o	ZS	.851 (.97/.76)	.546 (.64/.48)	.528 (.61/.47)	.597 (.61/.59)	.490 (.55/.44)	.681 (.66/.70)	.400 (.67/.29)
	FS	.873 (.97/.79)	.692 (.79/.62)	.695 (.74/.66)	.564 (.56/.57)	.269 (.50/.18)	.641 (.59/.70)	.133 (1.0/.07)
GPT-5-mini	ZS	.774 (.79/.76)	.674 (.75/.61)	.625 (.67/.59)	.234 (.80/.14)	.612 (.65/.58)	.740 (.80/.69)	.429 (.60/.33)
	FS	.881 (.94/.83)	.670 (.74/.61)	.658 (.68/.64)	.553 (.63/.49)	.235 (.46/.16)	.705 (.69/.72)	.000

Table 6: Baseline performance of proprietary models. Metrics: F1 (Precision/Recall). FM = *FamilyMember*, Obs. = *Observation*, Rel = FM-Obs Relation, Neg. = *Negated*, Living = *LivingStatus*, Side = *SideoftheFamily*. ZS = Zero-Shot, FS = Few-Shot.

subdataset, but increased for the *FHexchange-MTS* subdataset, with GPT-5-mini improving performance over GPT-4o with the provided examples. Overall, GPT-5-mini had greater zero-shot performance to extract *Observations*, which translated to stronger performance to relate *FamilyMembers* to *Observations*. Across both models, performance to extract age-related concepts was poor for *FHexchange-MTS*, missing all mentions of *Age*, *AgeofOnset*, and *AgeofDeath*. UMLS entity linking (CUI normalization) remained a difficult task across both subdatasets (best F1 = 0.373 for *FHexchange-KIT* with few-shot GPT-4o and F1 = 0.597 for *FHexchange-MTS* with zero-shot GPT-4o). In Appendix Table 13, we show three examples of differences between the models in extracting CUIs. GPT-5-mini was more conservative in assigning CUIs to *Observations* than GPT-4o. Few-shot prompting, where example CUI mappings were provided, is hypothesized to have increased the performance of GPT-5-mini to assign CUIs (0.115 to 0.255 for *FHexchange-KIT* and 0.234 to 0.553 for *FHexchange-MTS*).

As detailed in Appendix Tables 11 for *FHexchange-KIT* and 12 for *FHexchange-MTS*, the performance of Llama 3.3 was poor for all entities in comparison to the proprietary models. This was due to structural output errors in which the model did not correctly adhere to the JSON schema, despite correctly identifying entities within the reasoning field; however, few-shot prompting did improve the performance. This is a limitation, which may be addressed through additional prompt examples, fine-tuning, or parsing and normalization, and is an area for future work.

5 Discussion

The effect of collection methods on the characteristics and condition coverage of FHx data To enable more patient-centered care, there is a need for datasets derived from patient input. Patients report health data that is unique from their clinicians, providing additional context for a more holistic health profile (Castro et al., 2025). Data collection methods (surveys versus in-person medical history interviews) inform the characteristics and quality of the resulting dataset (Chen et al., 2015). Data from surveys with multiple choice options normalize the range of conditions and family members collected, due to the limitation of choices, whereas open-ended dialogs can lead to responses with more variance. Together, *FHexchange-MTS* and *FHexchange-KIT* provide a complementary view of modern patient-derived data for FHx, originating from their collection methods.

FHexchange-MTS is based on open-ended, free-form conversations about FHx. Not all dialogs contain specific prompting by doctors to patients, and in general, patients are tasked with spontaneously recalling relevant FHx. Because of this data collection method, we identify that *FHexchange-MTS* contains more colloquial vernacular, such as "sugar" for diabetes and "alcoholism" for alcohol use disorder. We provide mappings to the relevant CUI codes to standardize this language to computable clinical concepts. Additionally, the breadth of reported information was greater for *FHexchange-MTS* in comparison to *FHexchange-KIT* (13 UMLS Semantic Types captured versus eight Semantic Types).

In contrast, KIT users were prompted to recall their FHx for specific condition types and with

multiple choice options during their dialogs, and therefore, users relied more on option selection than free-text inputs. Because of this data collection method, the majority of responses within *FHexchange-KIT* were mapped to distinct options, leaving little variation in how reported conditions were expressed. When cross-mapping from UMLS codes to the ICD-10 clinical hierarchy, 39% of the unique concepts extracted from the open-ended *FHexchange-MTS* lacked a direct ICD-10 equivalent, compared to only 22% in the structured *FHexchange-KIT* survey data. This difference reinforces that open-ended dialogs yield a significantly wider variance of informal patient expressions, and structured dialogs may enable cleaner extraction into strict diagnostic hierarchies and computable formats to support automated risk assessments.

The potential for collection methods to inform what FHx data patients choose to disclose The open-ended dialogs in *FHexchange-MTS* show high enrichment for Circulatory System and Neoplasm-related disorders (12 and 18 unique codes respectively). The structured KIT dialogs show deeper representation across chronic condition systems like Digestive System (11 unique concepts, as compared to *FHexchange-MTS*'s three unique Digestive concepts). This may also suggest that asynchronous, comprehensive, structured methods of collecting FHx may be more effective to elicit both a broader (capturing a wider range of condition categories) and more specific (capturing finer grained condition language) array of FHx data than possible within limited conversations in a non-specialist setting. Asynchronous surveys conducted through chatbots may also lead to a higher willingness to share sensitive information when compared to face-to-face interactions with healthcare providers (Lucas et al., 2014). Supporting this notion, we found that two of the top five conditions represented within *FHexchange-KIT* were mental health related conditions, which patients may experience a stigma to report in traditional doctor-patient interactions (Nadarzynski et al., 2019). However, the populations and clinical settings represented in these datasets may also be distinct from one another, so future work should explore FHx data collection methods across the same populations and clinical contexts.

The utility of *FHexchange* to support the development of new clinical AI tools In the future, clinical AI tools will capture FHx across multiple modalities, from ambient scribes transcribing and documenting natural patient-doctor interactions to more flexible conversational agents (text and audio) providing medical history taking interviews (Hindelang et al., 2024). We contribute *FHexchange*, a unified FHx dialog dataset of both structured chatbot-user dialogs as well as unstructured human-human dialogs. The baseline evaluations of proprietary and open-weight models support the value of this corpus. While the proprietary language models accurately extracted basic family members from structured inputs, they did not extract temporal modifiers (ages) and other condition modifiers (*CauseofDeath*). Additionally, the UMLS CUI entity linking performance suggests that the field still needs robust, standardized mapping workflows to achieve computable clinical concepts for FHx. Error analysis revealed that the model translated colloquial patient expressions (e.g., "kidney cancer" and "fast heart beat") into formal medical terminology ("renal cancer" and "tachycardia"). In future work, we will consider these conceptually equivalent extractions as true positives; however, this highlights a tendency for the model to perform clinical normalization and not just exact extraction of the text. Additionally, our results suggest that reasoning models like GPT-5-mini may be conservative in linking clinical observations to standard terminologies that reduce the linking of both incorrect and correct mappings. These findings motivate the work to introduce additional domain knowledge, through the integration of ontologies or knowledge graphs, to ground the mapping of family history mentions to standard terminologies in future studies (Gao et al., 2023, 2025; Pan et al., 2024). Furthermore, future work could explore fine-tuning open-weight models to improve performance, as even few-shot prompts boosted Llama 3.3's ability to correctly extract FHx information (Park et al., 2023). *FHexchange* provides an evaluation set to assess emerging clinical AI tools across data collection modalities.

6 Conclusion

We are contributing *FHexchange*, consisting of annotated FHx resources for entity identification, relation extraction, and concept normalization from two distinct consumer-facing sources. The base data (for entity identification and relation extraction) will be available publicly for the clinical NLP research community from this data repository: www.

github.com/tirilab/FHexchange. By combining structured chatbot dialogs with unstructured human-to-human dialogs, we provide a comprehensive FHx dataset designed to benchmark the performance of emerging clinical AI systems to extract and normalize FHx data across multiple conversational modalities.

Limitations

First, the quality of patient-reported health information is characterized by patient recall. We did not annotate entities for patients' certainty of recall, but this consideration has been explored in previous annotation schema (Chen et al., 2015). Second, we did not include additional inter-annotator agreement metrics, such as an annotator confusion matrix. Third, the evaluation of the baseline is currently on strict token-level matches of *Observations*; therefore, this is currently underestimating the performance of the models. Fourth, the models were prompted for structured output by providing a structured Pydantic schema for extraction. This approach has shown to decrease performance and model reasoning (Tam et al., 2024), and we did not process Llama 3.3's structured outputs beyond the processing performed for the OpenAI models in order to extract the usable information. Fifth, baseline results did not include a clinical baseline. We only performed one model run for each document for each model, and the temperature of the models was not set to 0. Therefore, further testing of the consistency of the results is needed, as well as exploration of other models for this task. Sixth, we did not specifically annotate for strength of observations (e.g., indications that many family members experience a similar condition type or the perceived severity of certain observations). Seventh, for these datasets, we focus on English language dialogs only. Future work to support other languages is necessary. Lastly, UMLS CUIs do not capture all commonly used patient and/or consumer language for conditions. Previous work has been done to add consumer health vocabulary that was not explored in this work (Pham et al., 2019; Amith et al., 2019).

Ethical Considerations

All chatbot transcripts were anonymized to remove any identifying information and timestamps. We utilize the UMLS Metathesaurus to standardize medical observations via sciSpacy's entity linking module. We accessed the Metathesaurus (version 2022AB in scispacy after 2023 and version 2025AB through the API) through an individual license obtained for this work. Although the base annotations of FHx will be freely available for the research community, the UMLS CUI linkages will be provided through protected access that ensures each researcher abides by UMLS licensing requirements.

Acknowledgments

We would like to thank the genetic counselors and annotators who donated their time to make this work possible.

References

Muhammad Amith, Kirk Roberts, Cui Tao, Licong Cui, and Hua Xu. 2019. Ontology of Consumer Health Vocabulary: providing a formal and interoperable semantic resource for linking lay language and medical terminology. Proceedings. IEEE International Conference on Bioinformatics and Biomedicine, 2019:1177–1178.

Mahmoud Azab, Stephane Dadian, Vivi Nastase, Larry An, and Rada Mihalcea. 2019. Towards Extracting Medical Family History from Natural Language Interactions: A New Dataset and Baselines. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1255–1260, Hong Kong, China. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Robert Bill, Serguei Pakhomov, Elizabeth S. Chen, Tamara J. Winden, Elizabeth W. Carter, and Genevieve B. Melton. 2014. Automated Extraction of Family History Information from Clinical Notes. AMIA Annual Symposium Proceedings, 2014:1709–1717.

Pål H. Brekke, Taraka Rama, Ildikó Pilán, Øystein Nytrø, and Lilja Øvrelid. 2021. Synthetic data for annotation and extraction of family history information from clinical text. Journal of Biomedical Semantics, 12(1):11.

Victor M. Castro, Vivian S. Gainer, Danielle M. Crookes, Shawn N. Murphy, and Justin Manjourides. 2025. Comparing patient-reported symptoms and structured clinician documentation in electronic

health records. Journal of the American Medical Informatics Association: JAMIA, 32(9):1454–1461.

Elizabeth S Chen, Elizabeth W Carter, Tamara J Winden, Indra Neil Sarkar, Yan Wang, and Genevieve B Melton. 2015. Multi-source development of an integrated model for family health history. Journal of the American Medical Informatics Association : JAMIA, 22(e1):e67–e80.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, Alex Hall, and Victorien Plot. 2026. Pydantic Validation.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. Scientific Data, 9(1):313.

W. Gregory Feero, Mary Beth Bigley, and Kristin M. Brinner. 2008. New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. Journal of the American Medical Informatics Association : JAMIA, 15(6):723–728.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. DR.BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing. Journal of biomedical informatics, 138:104286.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. JMIR AI, 4:e58670.

Sergey Goryachev, Hyeoneui Kim, and Qing Zeng-Treitler. 2008. Identification and Extraction of Family History Information from Clinical Reports. AMIA Annual Symposium Proceedings, 2008:247–251.

Michael Hindelang, Sebastian Sitaru, and Alexander Zink. 2024. Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review. JMIR Medical Informatics, 12(1):e56628.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics, 33(1):159–174.

Neal Lewis, Daniel Gruhl, and Hui Yang. 2011. Dependency Parsing for Extracting Family History. In 2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology, pages 237–242.

Neal Lewis, Daniel Gruhl, Hui Yang, San Francisco, San Jose, and San Francisco. Extracting Family History Diagnoses From Clinical Texts.

Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37:94–100.

Microsoft. a. gpt-4o.

Microsoft. b. gpt-5-mini.

Microsoft. c. Llama-3-3-70B-Instruct.

Sungrim Moon, Liwei Wang, Xuan Chen, Nan Wang, Sheila M. Manemann, Nicholas B. Larson, Suzette J. Bielinski, and Hongfang Liu. 2022. Bridging the Granularity Gap in Family History Information Extracted from Clinical Narratives. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2022:795–804.

MTSamples. Medical Transcription Sample Reports & Examples | MTSamples.

Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digital Health, 5.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task, pages 319–327. Conference Name: Proceedings of the 18th BioNLP Workshop and Shared Task.

Michelle Hoang Nguyen, João Sedoc, and Casey Overby Taylor. 2024. Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis. Journal of Medical Internet Research, 26(1):e55164. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7):3580–3599.

Hyeryun Park, Jiye Son, Jeongwon Min, and Jinwook Choi. 2023. Selective UMLS knowledge infusion for biomedical question answering. Scientific Reports, 13(1):14214.

Quynh Pham, Gary Graham, Carme Carrion, Plinio P. Morita, Emily Seto, Jennifer N. Stinson, and Joseph A. Cafazzo. 2019. A Library of Analytic Indicators to Evaluate Effective Engagement with Consumer mHealth Apps for Chronic Conditions: Scoping Review. JMIR mHealth and uHealth, 7(1):e11941. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth.

Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, pages 111–121, Brussels, Belgium. Association for Computational Linguistics.

Majid Rastegar-Mojarad, Sijia Liu, Yanshan Wang, Naveed Afzal, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. 2018. BioCreative/OHNLP Challenge 2018. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '18, page 575, New York, NY, USA. Association for Computing Machinery.

Maciej Rybinski, Xiang Dai, Sonit Singh, Sarvnaz Karimi, and Anthony Nguyen. 2021. Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis. JMIR Medical Informatics, 9(4):e24020.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.

SFHERE Project. 2014. Sample sentences and annotations for family history.

SFHERE Team. 2017. Digital Library for Social and Family History-Extraction-Representation, and Evaluation Project.

Feichen Shen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. 2021. Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR medical informatics, 9(1):e24008.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes. arXiv preprint. ArXiv:2104.13498 [cs].

João Figueira Silva, João Rafael Almeida, and Sérgio Matos. 2020. Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach. JMIR Medical Informatics, 8(12):e22898.

Luca Soldaini and Nazli Goharian. QuickUMLS: a fast, unsupervised approach for medical concept extraction.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models. arXiv preprint. ArXiv:2408.02442 [cs].

The All of Us Research Program Investigators. 2019. The "All of Us" Research Program. The New England journal of medicine, 381(7):668–676.

U.S. National Library of Medicine. Unified Medical Language System (UMLS).

Liwei Wang, Huan He, Andrew Wen, Sungrim Moon, Sunyang Fu, Kevin J. Peterson, Xuguang Ai, Sijia Liu, Ramakanth Kavuluru, and Hongfang Liu. 2023. Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers–Assisted Sublanguage Analysis. JMIR Medical Informatics, 11(1):e48072.

Robert S. Wildin, Donna J. Messersmith, and Elisa J. F. Houwink. 2021. Modernizing family health history: achievable strategies to reduce implementation gaps. Journal of Community Genetics, 12(3):493–496.

Kecheng Zhan, Weihua Peng, Ying Xiong, Huhao Fu, Qingcai Chen, Xiaolong Wang, and Buzhou Tang. 2021. Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study. JMIR Medical Informatics, 9(4):e23587.

L. Zhou, Y. Lu, C.J. Vitale, P.L. Mar, F. Chang, N. Dhopeshwarkar, and R.A. Rocha. 2014. Representation of Information about Family Relatives as Structured Data in Electronic Health Records. Applied Clinical Informatics, 5(2):349–367.

A Annotation Guidelines

Objective. The goal of this protocol is to annotate family history details in clinical conversational narratives to support the development and evaluation of family history extraction tools.

A.1 Excel annotation sheet format

Annotations are recorded in an Excel sheet.

Column	Header (Excel)	Description
A	Document ID	Document identifier from the note header. Used to link annotations to the source note.
B	FamilyMember OR Observation	Top-level entity for the row: FamilyMember, Observation, or NaN (if no family history mentioned in the note).
C	FamilyMember Type	The relative type when Column B = FamilyMember. May also contain numbering to disambiguate multiple relatives of the same type (e.g., "uncle #2").
D	LivingStatus OR SideoftheFamily OR Observation OR Age	Secondary entity linked to the FamilyMember when Column B = FamilyMember.
E	LivingStatus OR SideoftheFamily	The value for Column D (Alive/Dead or Maternal/Paternal).
F	Description of Previous Entities (i.e., observation description, age of onset. . .)	Free text field for the condition name and relevant details REQUIRED for all Observation rows.
G	AgeofOnset or AgeofDeath	The value (ranges allowed) of age of onset or age of death from an Observation for a FamilyMember
H	Negated	True if the Observation is not present
I	CauseofDeath	True if the Observation is the cause of death of the FamilyMember

Table 7: Excel columns used for annotation.

Each row corresponds to at least one annotated *Observation*, with columns used to link (i) a document identifier, (ii) an optional family member mention, and (iii) one or more associated attributes and observations. The annotation procedure is as follows.

A.2 Annotation workflow

Step 1: Column A - Document identifier. In Column A, record the Document ID from the header of each note. This identifier links all annotation rows to the source document.

Step 2: Column B - Select the top-level entity. In Column B, select the major entity type to be described in the row: *Observation* or *FamilyMember*. An *Observation* is a family-history-relevant condition mention that does not specify an associated family member, for example:

> "Patient has no family history of atrial fibrillation."

If a specific family member is mentioned, select *FamilyMember*. If there is no family history mentioned anywhere in the note, use the *NaN* dropdown option.

Step 3: Column C - Populate FamilyMember type. If Column B is *Observation*, specify the observation in Column F (Description of Previous Entities).

If Column B is *FamilyMember*, specify the family member mentioned in Column C (FamilyMember Type), which is suggested to be one of the following: Great-grandfather, great-grandmother, grandmother, grandfather, father, mother, aunt, uncle, cousin, sister, brother, daughter, son, grandson, granddaughter, sibling, niece, or nephew, child, sibling. Please enter other relationships as needed, if they are present in the text. If multiple family members of the same type are mentioned (e.g., multiple uncles), add numbering in Column E (e.g., "uncle #2").

Step 4: Column D - Add associated rows for each FamilyMember-Observation pair OR just Observation. If Column B is *FamilyMember*, capture three additional entity types as separate rows associated with the same FamilyMember Type: *LivingStatus*, *SideoftheFamily*, and *Observation*. Each associated unique Observation entity for a given FamilyMember-Observation pair should be entered on its own row (e.g., if a sister is alive and has two conditions, create two *Observation* rows for a FamilyMember of Sister and record LivingStatus in both rows). If general good health is mentioned, it may be recorded as an Observation. If *SideoftheFamily* is not mentioned (e.g., "two of my aunts had cancer" without maternal/paternal), do not add *SideoftheFamily* information.

Notes may include abbreviations for family members. Pronouns are resolved to the previously mentioned family member when applicable. For example:

> "Her mother had coronary artery disease and hypertension. She also died from sudden cardiac death."

Here, *She* refers to *mother*.

Step 5: Fill Column D/E for LivingStatus or SideoftheFamily. If Column D is *LivingStatus*, Column E should be *Dead* or *Alive*.

LivingStatus may be inferred from context when not explicitly stated; for example:

> (1) "cousin has stomach cancer"

→ LivingStatus = Alive;

(2) “cousin undergoing treatment for stomach cancer”

→ LivingStatus = Alive;

(3) “cousin passed away from stomach cancer”

→ LivingStatus = Dead.

If Column D is *SideoftheFamily*, Column E should be *Paternal* or *Maternal*.

Step 6: Record Observation details in Column F. For any *Observation* (whether selected in Column B or Column D), Column F (Description of Previous Entities) is REQUIRED. Use this free text field to specify the condition name and any relevant details. For all entity types, Column F (Description of Previous Entities) is used to record details. Distinguish missing/unknown information from explicitly absent information: *Observation* – Unknown, indicates missing information or uncertainty (e.g., “I don’t know”); *Observation* – None, indicates not relevant/not pertinent/not significant when explicitly stated or asked about.

Step 7: Columns G, H, and I - Modifiers, Negated, and Cause of Death Column G (AgeofOnset or AgeofDeath): Record the numeric value (or range) of the age of onset or age of death related to an Observation.

Column H (Negated): For negative or absent family history (e.g., "No history of breast cancer"), record the relevant FamilyMember (if present; otherwise select Observation). Specify the condition in Column F, and mark Column H as True. Distinguish missing/unknown information from explicitly absent information: Observation – Unknown indicates missing information or uncertainty (e.g., "I don’t know"); Observation – None indicates not relevant/not pertinent/not significant when explicitly stated or asked about.

Column I (CauseofDeath): If an Observation is explicitly stated as the cause of the family member’s death, mark Column I as True. If an *Observation* can be inferred as the cause of death based off the phrased order (i.e. mention of death and mention of condition within the sentence or following sentence), mark Column I as True. For example:

"Well, doctor, he passed away. He had mesothelioma."

→ Column I = True (for the mesothelioma *Observation*).

A.3 Annotation Example

Excerpt from document 945

Doctor: Can you tell me more about your family history, ma’am? How is your father doing?

Patient: Well, doctor, my father passed away from leukemia a while back.

Column	Field	Value
A	doc_id	945
B	FamilyMember	Father
C	AgeofOnset	–
D	Observation	leukemia
E	SideoftheFamily	–
F	LivingStatus	Dead
G	Age	–
H	AgeofDeath	–
I	CauseofDeath	TRUE
K	Negated	–

Table 8: Example annotation from document 945.

B Error analysis between annotators

The following section details error analysis for annotations in both datasets.

Table 9 details the disagreements found after the first pass of annotations (before adjudication). Annotators are denoted by A and B. Semantic disagreement examples can be seen in Table 10.

Category	KIT	MTS
Exact/Boundary Match	629 (87.6%)	181 (67.8%)
Semantic Disagreement	31 (4.3%)	3 (1.1%)
Missing in B	32 (4.5%)	48 (18.0%)
Extra in B	26 (3.6%)	35 (13.1%)

Table 9: Annotator agreement breakdown for *Observations* and *FamilyMembers*. Missing in B means the mention was found extra within A, and extra in B signifies the mention was missing from B. KIT: n=718, MTS: n=267.

The examples in Table 10 were considered errors under a strict match policy but were considered lenient matches. In the final dataset, differences were adjudicated through discussion and removing extraneous information:

C Model prompt and FHx schema

C.1 Schema Definition

We define two Pydantic classes for structured family history extraction. The `FHx_Object` class represents a single family history entry with the fol-

#	Ann. A	Ann. B	Final
1	**FM:** Uncle **Obs:** had polio, leg deformities	**FM:** Uncle **Obs:** polio, leg deformities	**FM:** Uncle **Obs:** polio, leg deformities
2	**FM:** Father **LS:** Dead **Age:** 40s	**FM:** Father **LS:** Dead **Age:** died in 40s	**FM:** Father **LS:** Dead **Age:** 40s
3	**FM:** Brother **LS:** Alive unknown	**FM:** Brother **LS:** Alive	**FM:** Brother **LS:** Alive

Table 10: Lenient matches with final annotations. FM = *FamilyMember*, Obs = *Observation*, LS = *LivingStatus*

lowing fields: `reasoning` (string), `FamilyMember` (optional string), `AgeofOnset` (optional integer), `Observation` (optional string), `SideoftheFamily` (optional literal: "Maternal", "Paternal", or "Unknown"), `LivingStatus` (optional literal: "Alive", "Dead", or "Unknown"), `Age` (optional integer), `AgeofDeath` (optional integer), `CauseofDeath` (optional boolean), `CUI` (optional string), and `Negated` (optional boolean). The `FamilyHistory` class contains a list of `FHx_Object` instances called `full_history`.

C.2 Base System Prompt

Role: You are an assistant that is reviewing various texts to extract the relevant family health history information for several different condition groups.

Scope: Proband's relatives only. Exclude the proband's own conditions and any partner/spouse and their relatives.

Output: One `FamilyHistory` object with a list of `FHx_Objects`. Each `FHx_Object` may represent:

- A condition (`Observation`) linked to a specific family member, or
- A condition with no specific member (`FamilyMember` = null), or
- A family member record without a condition.

Fill null for any field not explicitly stated. Include negative and unknown family history.

FamilyMember relation fields: `LivingStatus` (alive/dead/unknown), `Age` (integer if explicit), `AgeOfDeath` (integer if explicit), `SideOfFamily` only for second-/third-degree relatives when explicitly stated.

Observation relation fields: `CUI` (UMLS target concept; null if uncertain), `AgeOfOnset` (integer if explicit), `CauseOfDeath` = true only if the text explicitly says the condition caused death; otherwise false.

Negated field: boolean; set to true ONLY if the text explicitly states a negative family history or that the member is healthy.

Provide concise rationale for extracted `FHx_Object` and reasons behind CUI mapping in the `reasoning` field.

Rules:

1. Extract only explicitly stated facts; otherwise use null. Do not infer ages, sex, side, or subtypes.
2. **Negateds:** If a text explicitly states a negative family history (e.g., "no history of cancer" or "brother is healthy"), extract the base condition in the `Observation` field (e.g., "cancer") and set the `Negated` field to true. For general healthy statements, set `Observation` to "any conditions" and `Negated` to true. Do not write the word "no" in the `Observation` field.
3. De-duplicate by member + condition (CUI) + status. Have a unique `FHx_Object` per condition per member.

C.3 Differences between Zero-shot and Few-shot approaches

The few-shot in-context learning approach provided one example from both *FHexchange* subdatasets. More details regarding the examples can be found in the dataset repository on Github.

D Baseline Experiment Results and Error Analysis

Model	Setting	FM	Obs.	FM-Obs.	CUI	Obs.-CUI
GPT-4o	ZS	.994 (.99/1.0)	.578 (.53/.64)	.632 (.66/.61)	.354 (.36/.35)	.229 (.23/.23)
	FS	.986 (.98/.99)	.638 (.58/.71)	.702 (.71/.69)	.373 (.37/.38)	.258 (.23/.29)
.000						
GPT-5-mini	ZS	.972 (.97/.97)	.726 (.66/.81)	.824 (.85/.80)	.115 (.48/.07)	.100 (.42/.06)
	FS	.891 (.97/.82)	.649 (.63/.67)	.711 (.79/.65)	.255 (.44/.18)	.189 (.37/.13)
Llama 3.3	ZS	.077 (1.0/.04)	.000	.000	.000	.000
	FS	.201 (.93/.11)	.101 (.83/.05)	.100 (.86/.05)	.107 (.11/.11)	.010 (.08/.01)

Table 11: Complete performance on *FHexchange-KIT*. Values: F1 (Precision/Recall). FM = *FamilyMember*, Obs. = *Observation*, ZS = Zero-Shot, FS = Few-Shot.

Model	Set	FM	Obs.	FM-Obs.	CUI	Obs.-CUI	Neg.	Obs.-Neg.	Living	FM-LS	Side	FM-Side
GPT-4o	ZS	.851 (.97/.76)	.546 (.64/.48)	.528 (.61/.47)	.597 (.61/.59)	.307 (.36/.27)	.983 (.98/.99)	.490 (.55/.44)	.681 (.66/.70)	.617 (.66/.58)	.400 (.67/.29)	.500 (1.0/.33)
	FS	.873 (.97/.79)	.692 (.79/.62)	.695 (.74/.66)	.564 (.56/.57)	.386 (.42/.36)	.982 (.97/.99)	.269 (.50/.18)	.641 (.59/.70)	.607 (.61/.60)	.133 (1.0/.07)	.105 (1.0/.06)
GPT-5-mini	ZS	.774 (.79/.76)	.674 (.75/.61)	.625 (.67/.59)	.234 (.80/.14)	.106 (.47/.06)	.984 (.98/.99)	.612 (.65/.58)	.740 (.80/.69)	.613 (.70/.55)	.429 (.60/.33)	.345 (.46/.28)
	FS	.881 (.94/.83)	.670 (.74/.61)	.658 (.68/.64)	.553 (.63/.49)	.371 (.49/.30)	.980 (.97/.99)	.235 (.46/.16)	.705 (.69/.72)	.680 (.71/.65)	.000	.000
Llama 3.3	ZS	.043 (.83/.02)	.000	.000	.000	.000	1.00 (1.0/1.0)	.000	.889 (1.0/.80)	.038 (1.0/.02)	.000	.000
	FS	.568 (.93/.41)	.468 (.78/.33)	.478 (.74/.35)	.405 (.39/.42)	.188 (.30/.14)	.992 (.98/1.0)	.169 (.50/.10)	.688 (.64/.75)	.452 (.64/.35)	.222 (.50/.14)	.174 (.40/.11)

Table 12: Complete performance on *FHexchange-MTS*. Values: F1 (Precision/Recall). FM = *FamilyMember*, Obs = *Observation*, Neg. = *Negated*, LS = *LivingStatus*. ZS = Zero-Shot, FS = Few-Shot. Age-related attributes were 0.000 across all models and omitted.

Observation	GPT-4o Reasoning	GPT-5-mini Reasoning
Testicular cancer (Father)	"User confirmed the father has been diagnosed with testicular cancer." CUI field: C0039590	"Father was indicated to have had a cancer described by the user as 'testicular'... CUI set to null because a precise UMLS mapping was not provided/confirmed." CUI field: null
Cataracts (Mother)	"Based on the selection indicating the mother has had hearing and/or eye conditions, further specifying cataracts." CUI field: C0007169*	"The user explicitly selected 'Mother' as having a hearing/eye condition... CUI is set to null because a specific UMLS concept mapping was not provided and is uncertain." CUI field: null
Heart attack (Father)	"User confirmed the father has experienced a heart attack." CUI field: C0027051	"Father was selected as having a heart and blood condition... CUI set to null because a precise UMLS mapping was not provided/-confirmed." CUI field: null

Table 13: Analysis of model behavior through schema *reasoning* field with CUI. GPT-4o assigns UMLS codes to *Observations* at a higher rate than GPT-5-mini (81.3% versus 10.4% assignment rate, respectively), but contains erroneous mappings and hallucinations*. For example, C0007169 is not a valid CUI. Few-shot prompting increases GPT-5-mini's CUI F1 from 0.115 to 0.255 (*FHexchange-KIT*) and 0.234 to 0.553 (*FHexchange-MTS*).

Doc	Predicted	Gold
FHex-MTS		
735	aunt, type 2 diabetes mellitus	aunt, type 2 diabetes
1161	grandmother, tachycardia	grandmother, fast heart beat
945	mother, kidney failure	mother, renal failure
FHex-KIT		
75	grandparent, irritable bowel syndrome	grandparent, irritable bowel syndrome (ibs)
86	father, leukemia	father, blood cancer
86	grandparent, renal cancer	grandparent, kidney cancer
57	sibling, substance abuse	sibling, drug use disorder

Table 14: *Observation* extraction errors showing semantic mismatches between model predictions and gold annotations.

Forgotten Words: Benchmarking NeoBERT for Dementia Detection in Low-Resource Conversational Filipino and English Speech

Rez Samantha Z. Floresca[1*], **Edric Castel C. Hao**[2†‡], **Hannah Grachiella Buñales**[1], **Chelsea Dominique E. Temprosa**[1], **Georgianna Z. Reyes**[1], **Kervin Gabriel L. Chua**[1]
[1]Ateneo de Manila Senior High School [2]Analog Devices, Inc.
rez.samantha.floresca@student.ateneo.edu

Abstract

Dementia detection from spontaneous speech offers a scalable approach to cognitive screening, yet NLP systems remain predominantly English-centric. This limitation is especially acute in the Philippines, where Filipino–English code-switching is pervasive and no prior work has addressed NLP-based dementia detection. We present the first systematic evaluation of transformer-based dementia detection in Filipino speech and the first assessment of NeoBERT in a clinical NLP setting. To separate language from domain effects, we construct a parallel bilingual dataset of 4,000 DementiaBank-derived transcripts, with Filipino translations produced manually to preserve discourse-level markers of cognitive decline. We evaluate five model families, TF-IDF + LogReg, BERT, NeoBERT, XLM-R, and RoBERTa-Tagalog, under monolingual, zero-shot cross-lingual, and bilingual fine-tuning settings. We find that in-domain performance does not transfer across languages, with English-trained BERT dropping to Macro-F1 = 0.455 on Filipino, and that architectural modernization alone does not improve robustness. Bilingual fine-tuning, however, eliminates cross-lingual degradation across all transformer models, converging to Macro-F1 = 0.969–0.973. These results suggest that multilingual clinical NLP performance is driven primarily by linguistic coverage during training rather than model scale or architecture.

1 Introduction

Dementia detection from speech is a promising direction for scalable cognitive screening, as linguistic and discourse-level patterns often reflect early cognitive decline (Peled-Cohen and Reichart, 2025). However, most NLP systems for this task are developed and evaluated primarily on English-language data, limiting their robustness in multilingual and low-resource settings. This limitation is particularly relevant in the Philippines, where everyday speech frequently involves Filipino–English code-switching and informal conversational structure. As a result, it remains unclear whether failures arise from cross-lingual transfer limitations or English-centric biases in model pretraining, motivating our focus on conversational transcripts to isolate discourse-level markers of cognitive decline.

Progress in Filipino NLP, while growing, has been largely confined to classification tasks on news articles and social media posts, including sentiment analysis (Villavicencio et al., 2021), named entity recognition (Miranda, 2023), and hate speech detection (Cruz and Cheng, 2022). Even recent evaluation benchmarks such as Batayan (Montalan et al., 2025) and FilBench (Miranda et al., 2025), while broader in scope, remain anchored to common written-text tasks such as reading comprehension, natural language inference, and generation. Critically, none of this work addresses naturalistic speech, clinical discourse, or cognitive diagnostic tasks. In particular, Tagalog–English code-switching, pervasive in informal and clinical Filipino communication (Herrera et al., 2022), is systematically absent from existing evaluations. As a result, it remains unknown whether transformer-based models can capture the discourse-level degradation characteristic of dementia in Filipino speech.

Additionally, transformer-based encoders have become the dominant paradigm in clinical NLP (Devlin et al., 2019; Pappagari et al., 2020), yet their application to dementia detection remains largely confined to architectural variants that differ only in pretraining data. ClinicalBERT (Alsentzer et al., 2019), AD-BERT (Mao et al.,

[*]Lead author.
[†]Provided research mentorship and guidance.
[‡]This work does not relate to the author's position at Analog Devices, Inc.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1029–1040
July 3-4, 2026 ©2026 Association for Computational Linguistics

2023), and mBERT (Pires et al., 2019) all inherit the original BERT design unchanged. This leaves open the question of how architectural modifications affect the extent to which models capture cognitively diagnostic linguistic properties relevant to dementia detection.

In this work, we study dementia detection under controlled bilingual conditions using conversational speech transcripts in English and Filipino. The full experimental setup is described in Section 3.

Our contributions are threefold: (1) We present the first evaluation of NLP-based dementia detection in Filipino speech, using conversational transcripts manually translated from English under controlled domain conditions. (2) We provide the first empirical characterization of NeoBERT's behavior under clinical and cross-lingual conditions. (3) We analyze cross-lingual transfer from English to Filipino under controlled conditions, benchmarking English-only models against multilingual (XLM-RoBERTa) and language-matched (RoBERTa-Tagalog) baselines to quantify the effect of English-centric pretraining under language shift. Our code is publicly available.[1]

2 Related Work

2.1 Linguistic Markers of Cognitive Decline

Spontaneous speech in dementia and mild cognitive impairment (MCI) exhibits reduced lexical diversity, increased disfluency, syntactic simplification, and degraded referential coherence, patterns that manifest measurably before clinical diagnosis (Fraser et al., 2016; Segkouli et al., 2025) and have been linked to early tau accumulation in language-relevant cortical regions (Marier et al., 2026). A recent systematic review confirms that these markers are consistently recoverable from text alone, with linguistic-only models reaching an average diagnostic accuracy of 83% across 51 NLP studies of dementia detection (Shankar et al., 2025).

However, existing work is heavily concentrated on monolingual English speech elicited under controlled clinical tasks, most commonly the Cookie Theft picture description task from DementiaBank (Becker et al., 1994). Bilingualism introduces competing forces: lifelong code-switching is associated with delayed dementia onset through enhanced cognitive reserve (Berkes and Bialystok, 2022), while the syntactic mixing and lexical borrowing characteristic of Taglish may alter the observable linguistic signals on which current models rely (Flores, 2019). This makes Filipino a theoretically significant yet entirely absent deployment context in the dementia NLP literature.

2.2 Transformer-Based Models in Clinical NLP

In dementia detection, most systems adapt BERT through domain-specific pretraining: ClinicalBERT (Alsentzer et al., 2019) adapts to clinical notes, AD-BERT (Mao et al., 2023) targets Alzheimer's progression prediction, and mBERT (Pires et al., 2019) extends to multiple languages. These models differ mainly in pretraining data while sharing the same architecture, leaving open the question of whether architectural changes improve cross-lingual clinical performance.

Representation pooling is another underexplored yet consequential design choice. Most existing systems default to `[CLS]`-token pooling, a choice optimized for the pretraining objective but not for aggregating task-relevant information across the full sequence. Reimers and Gurevych (2019) demonstrate that mean pooling over final hidden states produces more stable and transferable sentence representations, particularly under domain and language shift, motivating our adoption of mean pooling across all transformer-based models.

2.3 Cross-Lingual Dementia Detection

Cross-lingual transfer in dementia detection has been explored almost exclusively between typologically similar European languages. The ADReSS-M challenge (Luz et al., 2024), including work by Tamm et al. (Tamm et al., 2023) and (DeLoach, 2026), established English-to-Greek transfer as a benchmark setting. Greek, however, shares significant Indo-European typological properties with English, limiting the generalizability of these findings to more distant transfer settings such as English-to-Filipino. Studies on dementia detection in low-resource, non-European languages remain rare. Adhikari et al. (2022) applied NLP-based detection to English DementiaBank transcripts translated into Nepali, while Lu and Chen (2025) extended this

[1] https://github.com/rezsam09/Filipino-English-Dementia-Classification

line of work to Amis, an indigenous Taiwanese language, finding that augmenting a small dataset of 80 scripts with translated English transcripts improved both robustness and accuracy. Neither study examines cross-lingual transfer nor evaluates multilingual pretraining strategies. No work has examined dementia detection in a Southeast Asian language or in any language spoken by a population where bilingualism and code-switching are the norm rather than the exception. Filipino, spoken by over 100 million people and characterized by pervasive Filipino–English code-switching (Herrera et al., 2022), represents exactly this gap.

2.4 Multilingual and Language-Matched Pretraining

Multilingual pretraining has emerged as a common strategy for cross-lingual transfer in low-resource settings. XLM-RoBERTa (Conneau et al., 2020), trained on over 2.5 TB of filtered CommonCrawl data across 100 languages, substantially outperforms mBERT on low-resource cross-lingual benchmarks. However, even strong multilingual models do not fully capture the lexical and syntactic properties of Filipino: RoBERTa-Tagalog (Cruz and Cheng, 2022), pretrained on TLUnified, a large-scale, thematically diverse Filipino corpus, demonstrates consistent gains over XLM-RoBERTa on Filipino classification benchmarks, suggesting that language-matched pretraining captures aspects of Filipino that multilingual models do not adequately model. Whether this advantage transfers to a clinical domain, where training data is scarce and domain shift compounds language shift, remains entirely unknown.

2.5 Architectural Modernization of Encoder Models

NeoBERT (Breton et al., 2025) represents the most recent advance in encoder architecture, replacing absolute positional embeddings with Rotary Positional Embeddings (RoPE) (Su et al., 2021), adopting Pre-LayerNorm (Xiong et al., 2020) and RMSNorm (Zhang and Sennrich, 2019) for training stability, introducing SwiGLU in place of GELU activation (Shazeer, 2020), and optimizing depth-to-width ratio for parameter efficiency. Importantly, it remains exclusively English-pretrained on RefinedWeb, a corpus of 600B tokens, 18 times larger than RoBERTa's training data. Despite outperforming BERT-large, RoBERTa-large, and ModernBERT on MTEB under identical fine-tuning conditions (Breton et al., 2025), NeoBERT has not been evaluated in any clinical or cross-lingual setting, leaving open the question of whether its architectural advances translate into robustness under the combined domain and language shift characteristic of low-resource clinical NLP.

3 Experimental Design

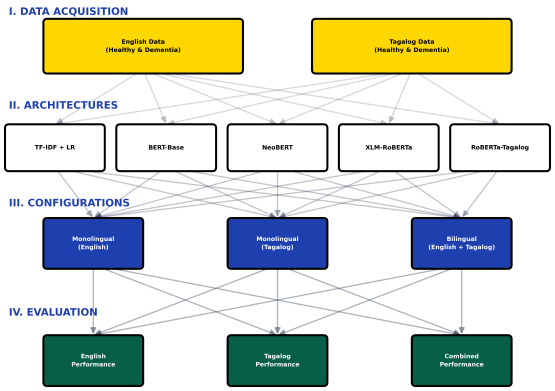


Figure 1: Overview of the experimental pipeline across datasets, model families, training configurations, and evaluation settings.

Performance improvements in NLP are often confounded by simultaneous changes in data, preprocessing, and model architecture (Sculley et al., 2015), a problem amplified in clinical NLP by small dataset sizes, domain sensitivity, and cross-linguistic variability (Fraser et al., 2016; Eyigoz et al., 2020). To mitigate this, all experiments use identical preprocessing pipelines with no external feature engineering beyond model-inherent representations, ensuring that observed performance differences reflect learned representations rather than preprocessing choices. The following subsections describe the dataset construction, preprocessing pipeline, model selection, and evaluation protocol.

3.1 Dataset Construction and Preprocessing

We construct a balanced bilingual dataset of 4,000 conversational transcripts for binary dementia classification, comprising 1,000 samples per class per language. All samples originate exclusively from DementiaBank (Becker et al., 1994), ensuring that any observed cross-lingual performance degradation reflects linguistic shift rather than

differences in topic distribution, discourse register, or clinical task structure.

3.1.1 English Data Sources

English dementia-positive and healthy control samples are drawn directly from DementiaBank (Becker et al., 1994), a widely used corpus in clinical speech research consisting of elicited speech from structured clinical tasks, most notably the Cookie Theft picture description task. This controlled elicitation procedure yields comparable discourse conditions across participants (Fraser et al., 2016), ensuring that class-level differences reflect genuine cognitive and linguistic variation rather than task-induced confounds.

3.1.2 Filipino Data Sources

Filipino samples are constructed through controlled manual translation of the complete set of 2,000 DementiaBank transcripts into Filipino, matching the English subset on clinical content, discourse structure, elicitation task, and class distribution. This ensures that cross-lingual evaluation is not confounded by asymmetric domain exposure. Manual translation is both necessary and principled given the absence of large-scale annotated clinical dementia corpora in Filipino (Mave et al., 2018). Translators are given explicit instructions to retain discourse-level markers of cognitive decline, including repetitions, hesitations, false starts, and syntactic degradation. Machine translation is avoided, since neural systems are known to normalize disfluent speech toward fluent output (Salesky et al., 2019), systematically erasing the features that distinguish dementia from healthy speech. All samples are anonymized prior to training in accordance with standard clinical data handling protocols.

3.1.3 Preprocessing Pipeline

All transcripts undergo a standardized preprocessing pipeline applying Unicode normalization, whitespace normalization, and lowercasing. Disfluencies, including filled pauses, repetitions, and hesitation markers, are retained throughout, as these are established linguistic correlates of cognitive impairment (Fraser et al., 2016) and primary classification signals in dementia detection. Stemming, lemmatization, and syntactic parsing are avoided, since such transformations risk normalizing disfluency structures and degrading the diagnostic signal available to downstream models. All input sequences are truncated to a maximum length of 128 tokens, consistent with the short, task-elicited structure of DementiaBank transcripts and with findings that neurocognitive markers are recoverable from short windows of speech (Fraser et al., 2016).

3.1.4 Models and Baselines

We evaluate five models across three pretraining paradigms. TF-IDF with Logistic Regression serves as a lexical baseline to assess whether surface-level token statistics suffice. BERT-base-uncased (Devlin et al., 2019) and NeoBERT (Breton et al., 2025) represent English-only pretraining, differing in architecture and corpus scale to isolate design capacity from language coverage. XLM-RoBERTa (Conneau et al., 2020) provides a 100-language multilingual baseline, while RoBERTa-Tagalog (Cruz and Cheng, 2022) evaluates language-matched pretraining on native Filipino text. These models decompose cross-lingual robustness along two orthogonal axes: architecture and language coverage. Because all models share an identical pipeline, cross-lingual performance variations reflect pretraining language exposure rather than fine-tuning differences.

Training Configurations: Models are trained separately for three setups: (1) English-only ($\mathcal{D}_{\text{EN}}$), (2) Filipino-only ($\mathcal{D}_{\text{TL}}$), and (3) Bilingual ($\mathcal{D}_{\text{EN}} \cup \mathcal{D}_{\text{TL}}$), enabling controlled analysis of monolingual capacity, cross-lingual transfer, and bilingual transfer.

3.2 TF-IDF with Logistic Regression

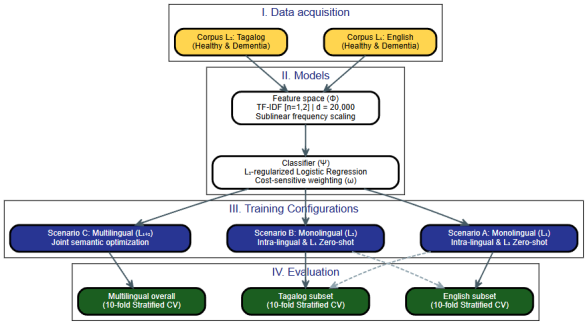


Figure 2: TF-IDF + Logistic Regression baseline training pipeline for evaluating lexical separability under cross-lingual and clinical domain shift conditions

We use TF-IDF with Logistic Regression as a lexical baseline to assess whether dementia-related

signals are captured by surface-level token statistics alone. TF-IDF produces sparse lexical representations by weighting tokens based on frequency statistics while ignoring word order and syntactic structure. Logistic Regression then learns a linear decision boundary over this high-dimensional feature space. This combination is widely used as an interpretable baseline in text classification tasks (Guo et al., 2020; Wang and Manning, 2012; Zhang et al., 2015). Performance of this model provides a reference for lexical separability: strong results suggest that class information is recoverable from surface-level features alone, whereas a weaker performance indicates the need for contextual modeling.

3.2.1 TF-IDF Representation

We construct unigram and bigram TF-IDF features ($n \in 1, 2$), capturing both lexical identity and short disfluency-related phrases. Bigrams are sufficient given the short, utterance-level structure of clinical transcripts and have been shown to yield consistent gains in text classification tasks (Wang and Manning, 2012). We apply standard TF-IDF preprocessing settings: (1) sublinear term-frequency scaling ($1 + \log(tf)$) to reduce the influence of repeated tokens common in disfluent speech; (2) a minimum document frequency of 2 (absolute count) to remove rare transcription noise; (3) a maximum document frequency of 0.95 to filter non-informative high-frequency tokens; and (4) a maximum vocabulary size of 20,000, consistent with standard practice in text classification (Zhang et al., 2015). These constraints ensure a stable lexical feature space that is comparable across languages and avoids overfitting in low-resource settings.

3.2.2 Logistic Regression Classifier

We use ℓ_2-regularized logistic regression due to its convex optimization properties and strong performance in high-dimensional sparse settings. Optimization is performed using the `liblinear` solver for efficient and deterministic convergence on sparse data, with a maximum of 2,000 iterations to ensure convergence.

3.3 Transformer-Based Models

We fine-tune four transformer-based models for binary dementia classification: BERT-base-uncased, NeoBERT, XLM-RoBERTa, and RoBERTa-Tagalog. Together, these models span three pretraining paradigms: English-only, massively multilingual, and language-matched, enabling a systematic decomposition of cross-lingual robustness along both architectural and pretraining dimensions.

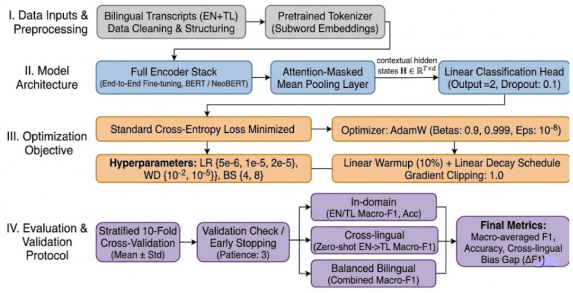


Figure 3: Transformer Models Fine-tuning Pipeline

3.3.1 Model Architecture and Fine-Tuning

All models follow a standard transformer-based classification pipeline. Given an input transcript x, tokens are first mapped to subword embeddings using a model-specific pretrained tokenizer, producing input IDs and attention masks. The resulting sequence is passed through the encoder, which is fine-tuned end-to-end without frozen layers. A single sentence representation is obtained via attention-masked mean pooling over the final hidden states:

$$h = \frac{\sum_{i=1}^{T} m_i H_i}{\sum_{i=1}^{T} m_i}$$

where m_i denotes the attention mask, ensuring that padding tokens do not contribute to the sentence representation. Unlike CLS pooling, which is optimized for pretraining objectives, masked mean pooling aggregates distributed token-level information and has been shown to improve robustness under domain shift (Reimers and Gurevych, 2019). The resulting representation is passed to a dropout layer ($p = 0.1$) followed by a linear classification head. The encoder is initialized from pretrained weights, while the classification head is randomly initialized.

3.3.2 Optimization Objective

We minimize standard cross-entropy loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y)} \log p_\theta(y \mid x)$$

Optimization is performed using AdamW (Loshchilov and Hutter, 2019), which decouples weight decay from gradient updates and improves generalization in transformer fine-tuning settings.

3.3.3 Hyperparameter Selection

Hyperparameters are selected via a systematic grid search designed to avoid loss divergence and unstable gradient updates on small clinical datasets. The search space for all BERT models includes: (1) learning rates $\{5 \times 10^{-6}, 6 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$ (2) weight decay $\{10^{-2}, 10^{-5}\}$ (3) batch sizes $\{4, 8\}$. These ranges are motivated by the constraints of low-resource clinical fine-tuning. Lower learning rates help prevent catastrophic forgetting of pretrained representations, while varying weight decay controls the balance between task-specific adaptation and robustness to noise introduced by manual transcription and translation. We use a stratified 70–15–15 train/validation/test split. The training set is used for model fitting, the validation set for early stopping, and the test set is used for hyperparameter selection. Small batch sizes $(4, 8)$ are chosen to improve generalization through gradient stochasticity (Masters and Luschi, 2018) while maintaining memory efficiency for fine-tuning NeoBERT. This also increases the number of parameter updates per epoch, which is beneficial in low-data regimes.

Table 1 reports the optimal hyperparameter configurations identified via grid search for each architecture and training regime. Across models, learning rates in the range of 1×10^{-5} to 3×10^{-5} consistently yield stable fine-tuning, with lower rates (1–2×10^{-5}) favored for architecturally complex or multilingual models such as XLM-RoBERTa and NeoBERT, likely reflecting greater sensitivity to catastrophic forgetting. NeoBERT consistently selects a batch size of 4 across all regimes, suggesting that higher gradient stochasticity benefits its fine-tuning under low-resource conditions. Weight decay varies across configurations, with stronger regularization (1×10^{-2}) generally preferred in monolingual English and Filipino settings, while smaller values (1×10^{-5}) are selected more frequently in bilingual and cross-lingual regimes, consistent with the need for greater representational flexibility under mixed-language supervision.

Architecture	Regime	Batch	LR	WD
BERT	English	8	2×10^{-5}	1×10^{-2}
BERT	Tagalog	4	3×10^{-5}	1×10^{-5}
BERT	Bilingual	4	3×10^{-5}	1×10^{-2}
XLM-RoBERTa	English	8	1×10^{-5}	1×10^{-5}
XLM-RoBERTa	Filipino	8	3×10^{-5}	1×10^{-2}
XLM-RoBERTa	Bilingual	8	2×10^{-5}	1×10^{-5}
RoBERTa-Tagalog	English	8	3×10^{-5}	1×10^{-2}
RoBERTa-Tagalog	Filipino	4	3×10^{-5}	1×10^{-2}
RoBERTa-Tagalog	Bilingual	8	3×10^{-5}	1×10^{-5}
NeoBERT	English	4	2×10^{-5}	1×10^{-5}
NeoBERT	Tagalog	4	3×10^{-5}	1×10^{-5}
NeoBERT	Bilingual	4	3×10^{-5}	1×10^{-5}

Table 1: Optimal hyperparameter configurations for transformers across all training regimes.

3.4 Training Procedure

Each configuration is trained independently from a fresh pretrained checkpoint to avoid parameter leakage across hyperparameter settings. Training is conducted under the following constraints: (1) a maximum of 10 epochs, which is sufficient for convergence in low-resource fine-tuning, and (2) full backpropagation over all model parameters. During hyperparameter search, early stopping with a patience of 3 is applied using validation macro-F1 to prevent overfitting. For final reported results, models are retrained without early stopping to ensure a consistent number of training epochs across experimental conditions.

3.4.1 Cross-Validation Protocol

Clinical NLP datasets are typically small and high-variance, making single train–test splits prone to unstable performance estimates. To evaluate robustness with respect to data variability, we perform stratified 10-fold cross-validation on the data for all models and training configurations. Final cross-validation performance is reported as the mean and standard deviation across folds. Importantly, cross-validation is not used for model selection in the final reported results. Instead, it is used to assess performance stability and sensitivity to data sampling.

3.4.2 Evaluation Protocol and Metrics

For each training configuration, we evaluate under three settings. In-domain evaluation trains and tests within the same language using stratified 10-fold cross-validation. Cross-lingual zero-shot evaluation trains on each fold of one language and evaluates on the full dataset of the other language, which is never seen during training. This ensures a consistent cross-lingual test condition across folds, with variance reflecting sensitivity to training data sampling rather than target-language variability. Bilingual evaluation trains on each fold of the combined corpus ($\mathcal{DEN} \cup \mathcal{DTL}$) and evaluates on the corresponding held-out test fold. Performance is reported on the full mixed-language test fold as well as separately on English-only

and Filipino-only subsets. Each configuration is trained independently. Cross-lingual generalization is assessed by training on one language and evaluating on the other in a zero-shot setting. The combined setting evaluates models on a mixed-language test set containing both English and Filipino samples. We report Accuracy and Macro-averaged F1-score as the primary evaluation metrics. Macro-F1 is prioritized due to its equal weighting of the Healthy and Dementia classes, making it more appropriate for imbalanced clinical classification tasks where minority-class performance is critical. To quantify robustness under language shift, we define the cross-lingual generalization gap as follows

$$\Delta_{\text{F1}} = \left|\text{F1}_{\text{in-domain}} - \text{F1}_{\text{cross-lingual}}\right| \quad (1)$$

where in-domain refers to evaluation on the same language as training (EN→EN or TL→TL), and cross-lingual refers to zero-shot evaluation on the unseen language (EN→TL or TL→EN). The combined setting is reported separately to assess performance under mixed-language conditions and is excluded from the cross-lingual generalization gap computation. In addition to overall performance, we report class-wise F1 scores and Dementia-class recall to explicitly evaluate clinical sensitivity. These metrics allow us to analyze whether performance degradation is driven by reduced detection of Dementia cases or by general classification instability.

4 Results and Discussion

		F1 ± std / Acc ± std			
Model	Train	EN	TL	Comb	Gap
TF-IDF + LR	EN	0.930 ± 0.013 / 0.930 ± 0.013	0.649 ± 0.008 / 0.684 ± 0.006	0.836 ± 0.005 / 0.839 ± 0.004	0.281
TF-IDF + LR	TL	0.640 ± 0.005 / 0.677 ± 0.003	0.971 ± 0.015 / 0.971 ± 0.015	0.825 ± 0.003 / 0.830 ± 0.003	0.331
TF-IDF + LR	EN+TL	0.932 ± 0.014 / 0.932 ± 0.014	0.976 ± 0.010 / 0.976 ± 0.010	0.954 ± 0.011 / 0.954 ± 0.011	0.044
BERT	EN	0.952 ± 0.014 / 0.952 ± 0.014	0.455 ± 0.012 / 0.561 ± 0.007	0.744 ± 0.008 / 0.755 ± 0.007	0.497
BERT	TL	0.705 ± 0.033 / 0.728 ± 0.027	0.981 ± 0.010 / 0.981 ± 0.010	0.855 ± 0.014 / 0.858 ± 0.013	0.276
BERT	EN+TL	0.954 ± 0.009 / 0.954 ± 0.009	0.984 ± 0.009 / 0.984 ± 0.009	0.969 ± 0.007 / 0.969 ± 0.007	0.030
XLM-RoBERTa	EN	0.948 ± 0.017 / 0.949 ± 0.017	0.936 ± 0.018 / 0.936 ± 0.017	0.942 ± 0.016 / 0.942 ± 0.016	0.013
XLM-RoBERTa	TL	0.825 ± 0.031 / 0.830 ± 0.029	0.986 ± 0.010 / 0.986 ± 0.010	0.905 ± 0.019 / 0.906 ± 0.019	0.161
XLM-RoBERTa	EN+TL	0.953 ± 0.010 / 0.954 ± 0.010	0.990 ± 0.007 / 0.990 ± 0.007	0.972 ± 0.006 / 0.972 ± 0.006	0.037
RoBERTa-Tagalog	EN	0.951 ± 0.014 / 0.951 ± 0.014	0.934 ± 0.005 / 0.934 ± 0.005	0.942 ± 0.015 / 0.943 ± 0.015	0.017
RoBERTa-Tagalog	TL	0.769 ± 0.045 / 0.781 ± 0.038	0.987 ± 0.006 / 0.987 ± 0.006	0.882 ± 0.018 / 0.884 ± 0.017	0.218
RoBERTa-Tagalog	EN+TL	0.958 ± 0.010 / 0.958 ± 0.010	0.988 ± 0.007 / 0.988 ± 0.007	0.973 ± 0.006 / 0.973 ± 0.006	0.030
NeoBERT	EN	0.952 ± 0.013 / 0.952 ± 0.013	0.617 ± 0.109 / 0.667 ± 0.084	0.802 ± 0.045 / 0.808 ± 0.042	0.335
NeoBERT	TL	0.757 ± 0.060 / 0.772 ± 0.051	0.979 ± 0.011 / 0.979 ± 0.011	0.878 ± 0.031 / 0.880 ± 0.030	0.222
NeoBERT	EN+TL	0.956 ± 0.015 / 0.956 ± 0.015	0.983 ± 0.009 / 0.983 ± 0.009	0.970 ± 0.007 / 0.970 ± 0.007	0.027

Table 2: Performance across models and training settings. Cells report F1 ± std / Acc ± std. Cross-lingual gap: $\Delta F1 = |F1_{\text{in-domain}} - F1_{\text{cross-lingual}}|$

Table 2 reports Macro-F1 and Accuracy across all models and training configurations. Four findings emerge consistently.

4.1 Cross-Lingual Failure Persists Even Under Controlled Parallel Data

All models achieve strong in-domain performance but degrade substantially under language transfer. English-trained BERT achieves $F1 = 0.952$ on English but falls to 0.455 on Filipino ($\Delta = 0.497$), while Filipino-trained BERT achieves $F1 = 0.981$ on Filipino but drops to 0.705 on English ($\Delta = 0.276$). The asymmetry is informative: transfer into English is consistently easier than transfer into Filipino, even when the model is fine-tuned entirely on Filipino, suggesting that English remains a stronger prior in the representation space due to pretraining exposure. Fine-tuning on Filipino modifies model behavior but does not fully overwrite the English-centric geometry learned during pretraining. These results cannot be explained by domain shift, since the Filipino corpus was constructed through controlled manual translation of the same source transcripts. The degradation, therefore, isolates language transfer rather than demographic or collection confounds. The failure appears representational: models learn dementia-related patterns within a language but fail to align those patterns robustly across languages. The lexical baseline supports this interpretation. TF-IDF + LR achieves $F1 = 0.649$ on Filipino when trained on English and $F1 = 0.640$ on English when trained on Filipino, suggesting that some dementia-related surface markers survive translation at the lexical level. The gap relative to transformer-based models indicates that lexical overlap alone is insufficient for stable cross-lingual generalization.

4.2 NeoBERT Improves Monolingual Fidelity While Increasing Transfer Instability

Relative to BERT, NeoBERT achieves comparable monolingual performance but does not improve cross-lingual robustness. Despite matching BERT's in-domain performance (F1=0.952), English-trained NeoBERT degrades substantially on Filipino (F1=0.617) and exhibits the highest variance across any experiment ($\sigma = 0.109$), suggesting that architectural modernization forms tighter English-side decision boundaries that improve in-domain fidelity while reducing tolerance to linguistic variation. Filipino-trained NeoBERT transfers better into English than BERT (F1=0.757 vs. 0.705) but retains high instability in dementia recall (Table 3). Taken together,

architectural improvements alone do not produce language-invariant clinical representations, and stronger monolingual specialization appears to trade off against transfer stability.

4.3 Transfer Behavior of Multilingual and Language-Matched Pretraining

XLM-RoBERTa and RoBERTa-Tagalog exhibit substantially smaller transfer gaps than English-only encoders, but through different mechanisms. English-trained XLM-RoBERTa achieves $F1 = 0.936$ on Filipino ($\Delta = 0.013$), the smallest gap observed under monolingual training, consistent with multilingual pretraining constructing a shared representation space that allows English fine-tuning to map naturally onto Filipino inputs. However, transfer is asymmetric: XLM-RoBERTa transfers substantially better from English to Filipino than the reverse ($\Delta = 0.013$ $vs.$ 0.161), suggesting that English occupies a disproportionately larger share of the pretraining corpus, producing a representation space calibrated more precisely around English structure. RoBERTa-Tagalog is more surprising. Despite being pretrained exclusively on Filipino, it achieves nearly identical English $\rightarrow$ Filipino transfer as XLM-RoBERTa ($F1 = 0.934$, $\Delta = 0.017$). We attribute this to the linguistic composition of Filipino itself: conversational Filipino contains extensive English lexical borrowing, code-switching, and Anglicized syntactic structure, meaning a model pretrained deeply on Filipino encounters substantial embedded English structure even without explicit multilingual supervision. The reverse direction supports this interpretation, as RoBERTa-Tagalog degrades more sharply when English becomes the target ($\Delta = 0.218$), suggesting that embedded English structure is sufficient for English-to-Filipino transfer but weaker than explicit multilingual supervision for reconstructing robust English-side representations after Filipino fine-tuning.

4.4 Bilingual Training Eliminates Most Cross-Lingual Bias

Bilingual fine-tuning substantially reduces cross-lingual degradation across all architectures, with combined-set Macro-F1 converging to 0.969–0.973 and cross-lingual gaps shrinking to between 0.027 and 0.037. Models that behaved very differently under monolingual transfer become nearly indistinguishable after bilingual supervision. NeoBERT, which exhibited the highest instability under English-only transfer, stabilizes immediately and performs comparably to XLM-RoBERTa and RoBERTa-Tagalog, suggesting that the primary bottleneck is not architectural capacity but representational alignment. The models already possess sufficient capacity to encode dementia-related linguistic structure across both languages; what differs is whether the optimization signal encourages those structures to occupy compatible regions of the embedding space. Multilingual and language-matched models retain a small but consistent advantage, with XLM-RoBERTa and RoBERTa-Tagalog achieving the highest combined-set performance (0.972–0.973), suggesting that pretraining language coverage provides a more favorable initialization for bilingual adaptation. Bilingual TF-IDF + LR substantially reduces the transfer gap ($\Delta = 0.044$) but remains below all transformer models, confirming that surface lexical statistics capture part of the dementia signal but are insufficient to construct stable cross-lingual abstractions.

		Healthy			Dementia			Dem. Recall		
Model	Train	EN	TL	Comb	EN	TL	Comb	EN	TL	Comb
TF-IDF+LR	EN	0.934	0.990	0.807	0.926	0.989	0.853	0.879	0.983	0.968
TF-IDF+LR	TL	0.955	0.971	0.855	0.952	0.970	0.795	0.931	0.952	0.660
TF-IDF+LR	EN+TL	0.972	0.976	0.956	0.971	0.974	0.952	0.957	0.950	0.922
BERT	EN	0.216	0.953	0.690	0.695	0.950	0.798	1.000	0.931	0.966
BERT	TL	0.981	0.786	0.875	0.981	0.624	0.836	0.979	0.455	0.724
BERT	EN+TL	0.984	0.955	0.969	0.984	0.954	0.969	0.982	0.940	0.961
XLM-RoBERTa	EN	0.935	0.950	0.942	0.936	0.947	0.941	0.937	0.920	0.928
XLM-RoBERTa	TL	0.986	0.855	0.914	0.985	0.794	0.896	0.977	0.661	0.815
XLM-RoBERTa	EN+TL	0.990	0.955	0.972	0.989	0.952	0.971	0.983	0.931	0.957
RoBERTa-Tagalog	EN	0.935	0.952	0.943	0.934	0.950	0.942	0.923	0.928	0.925
RoBERTa-Tagalog	TL	0.987	0.821	0.896	0.987	0.717	0.869	0.981	0.563	0.772
RoBERTa-Tagalog	EN+TL	0.988	0.959	0.973	0.987	0.957	0.972	0.986	0.939	0.963
NeoBERT	EN	0.504	0.952	0.774	0.729	0.951	0.830	0.915	0.939	0.927
NeoBERT	TL	0.979	0.816	0.893	0.978	0.699	0.863	0.969	0.544	0.766
NeoBERT	EN+TL	0.983	0.956	0.970	0.983	0.955	0.969	0.985	0.938	0.962

Table 3: Class-level performance across models and training configurations. Cells report F1; Healthy/Dementia are class F1; Dem. Recall is AD recall.

4.5 Clinical Implications Under Language Shift

Table 3 shows that aggregate Macro-F1 partially obscures clinically important failure modes under language transfer. English-trained BERT maintains high dementia recall on Filipino (0.931), but Healthy-class F1 collapses to 0.216, indicating that the model predicts most Filipino samples as dementia-positive rather than transferring meaningful discriminative structure. NeoBERT

exhibits a different pattern: dementia recall on Filipino averages 0.939 but with extremely large variance ($\sigma = 0.190$), reflecting inconsistent decision boundaries across fine-tuning folds rather than stable transfer. In contrast, English-trained XLM-RoBERTa and RoBERTa-Tagalog maintain balanced class F1 scores across languages with substantially lower variance, a property particularly important in screening settings where robustness across patient populations matters more than aggregate accuracy.

Bilingual training resolves most instabilities across all architectures, with all transformer models achieving dementia recall above 0.93 with consistently low variance, indicating substantially more stable decision boundaries than those learned under monolingual training.

4.6 Comparison with Classical Baselines

The lexical baseline shows moderate transfer in both directions: English-trained TF-IDF + LR achieves $F1 = 0.649$ on Filipino and 0.640 in the reverse direction, suggesting that some dementia-related surface markers are preserved across the translation boundary. Under English-only training, TF-IDF performs comparably to NeoBERT cross-lingually but remains substantially below XLM-RoBERTa and RoBERTa-Tagalog. In the bilingual setting, TF-IDF + LR reaches $F1 = 0.954$ while transformer models achieve 0.969–0.973, indicating that sparse lexical features are insufficient even under bilingual exposure. These results suggest that transformer gains stem not from stronger in-language modeling alone, but from learning contextual representations that transfer across languages. TF-IDF captures surface-level dementia markers but does not model the contextual interactions through which those markers manifest across languages, whereas multilingual and language-matched models better preserve these dependencies. Overall, TF-IDF retains partial transferability due to the structural persistence of dementia-related cues, but its instability relative to contextual models indicates that robust cross-lingual screening requires representation learning beyond lexical overlap.

5 Conclusion

Across all model families, strong in-domain performance does not translate to cross-lingual robustness, even when domain and clinical content are held constant through parallel dataset construction. This indicates that the primary barrier to multilingual clinical NLP is representational misalignment induced by language-specific pretraining rather than task complexity or domain shift. Architectural modernization alone, as shown by NeoBERT, does not yield consistent cross-lingual gains and may increase sensitivity to language shift through tighter monolingual specialization. Multilingual and language-matched pretraining provide more stable transfer, though through different mechanisms. Bilingual fine-tuning, however, largely eliminates cross-lingual performance gaps across all architectures, suggesting that linguistic coverage during task training is more influential than architectural choice in determining robustness under language shift.

These findings indicate that reliable clinical NLP in low-resource, code-switched settings depends less on model scaling or architectural modification and more on ensuring adequate linguistic coverage during training, with bilingual supervision emerging as the most effective strategy for stable and clinically consistent performance across languages.

6 Limitations

Several limitations of this study warrant acknowledgment. First, the Filipino dataset was constructed through controlled manual translation of English DementiaBank transcripts, as no large-scale native clinical dementia corpora exist for Filipino. Although translators preserved structural disfluencies, hesitations, and syntactic fragmentation, the translated text inherently reflects the conversational structure and semantic content of the original English source documents. The dataset's modest scale of 4,000 samples also restricts the generalizability of evaluation metrics and contributes to cross-validation variance. Acquiring larger, organically produced clinical speech from local Filipino patient cohorts in collaboration with geriatricians remains a priority for validating these findings in local clinical environments.

Second, this study focuses exclusively on text to isolate language-specific discourse patterns, setting aside acoustic and non-verbal features. In real-world clinical settings, vocal features such

as pitch variance, pause duration, and phonation rate provide independent diagnostic markers of neurodegeneration not captured in a text-only framework. Integrating speech encoders such as Wav2Vec 2.0 alongside transformer-based text models represents an important future direction.

Finally, while this evaluation establishes classification baselines across a range of encoder architectures, the mechanisms driving model decisions in multilingual contexts remain opaque. Clinical deployment demands transparency, and applying local feature attribution and interpretability methods is a necessary step before these screening workflows can achieve clinical trust and real-world deployment.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2022. Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques. *International Journal of Human-Computer Studies*, 160:102761.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

James T. Becker, François Boller, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

M. Berkes and Ellen Bialystok. 2022. Bilingualism as a contributor to cognitive reserve: What it can do and what it cannot do. *American Journal of Alzheimer's Disease & Other Dementias*, 37.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X. Morris, and Sarath Chandar. 2025. NeoBERT: A next-generation BERT. *Preprint*, arXiv:2502.19587.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Kylar A. DeLoach. 2026. Dementia detection in low-resource languages: Evaluating translation-assisted transfer learning for multilingual clinical assessment. Honors thesis, Mississippi State University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elif Eyigoz, Sachin Mathur, Mar Mishkind, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28:100583.

Eden Regala Flores. 2019. A study on patterns and functions of tagalog-english code switching in two oral discussions. *International Journal of TESOL Studies*.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Yuxia Guo, Xu Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cecile Paris, and Diego Mollá. 2020. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 104–112.

Megan Herrera, Ankit Aich, and Natalie Parde. 2022. TweetTaglish: A dataset for investigating Tagalog-English code-switching. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2592–2601, Marseille, France. European Language Resources Association.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA. OpenReview.net.

Chingching Lu and Shih-Wei Chen. 2025. Improving early detection of dementia in low-resource languages: A connected speech analysis case study in Amis language. *Alzheimer's & Dementia*, 20(Suppl 8):e095345.

Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and

Brian MacWhinney. 2024. An overview of the ADReSS-M signal processing grand challenge on multilingual Alzheimer's dementia recognition through spontaneous speech. *IEEE Open Journal of Signal Processing*, 5:738–749.

Chengsheng Mao, Jie Xu, Luke Rasmussen, Yikuan Li, Prakash Adekkanattu, Jennifer Pacheco, Borna Bonakdarpour, Robert Vassar, Li Shen, Guoqian Jiang, Fei Wang, Jyotishman Pathak, and Yuan Luo. 2023. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *Journal of Biomedical Informatics*, 144:104442.

Anna Marier, Jaime Fernández Arias, Étienne Aumont, Brandon J. Hall, Arthur C. Macedo, Nesrine Rahmouni, Gleb Bezgin, Paolo Vitali, Pedro Rosa-Neto, and Maxime Montembeault. 2026. Language deficits across PET-based Braak stages of tau accumulation in Alzheimer's disease. *Alzheimer's & Dementia*, 22(3):e71286.

Dominic Masters and Carlo Luschi. 2018. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Lester James V. Miranda. 2023. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. FilBench: Can LLMs understand and generate Filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah S. Flores, Michael T. Lopez II, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31239–31273, Vienna, Austria. Association for Computational Linguistics.

Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velázquez, and Najim Dehak. 2020. Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. In *Proceedings of Interspeech 2020*, pages 2177–2181.

Lotem Peled-Cohen and Roi Reichart. 2025. A systematic review of NLP for dementia: Tasks, datasets and opportunities. *Transactions of the Association for Computational Linguistics*, 13:1204–1244.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2786–2792.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 2503–2511. Curran Associates, Inc.

Sofia Segkouli, Mara Gkioka, Stylianos Kokkas, Konstantinos Votis, Sergi Valero, Andrea Miguel, Athos Antoniades, Emily Charalambous, and George Manias. 2025. Linguistic markers in spontaneous speech: Insights into subjective cognitive decline (review). *Healthcare*, 13(22):2888.

Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3:100205.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*.

Bastiaan Tamm, Rik Vandenberghe, and Hugo Van Hamme. 2023. Cross-lingual transfer learning for Alzheimer's detection from spontaneous speech. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–2.

Charlyn Nayve Villavicencio, Julio Jerison E. Macrohon, X. Alphonse Inbaraj, Jyh-Horng Jeng, and Jyh-Cheng Hsieh. 2021. Twitter sentiment analysis towards COVID-19 vaccines in the Philippines using naïve Bayes. volume 12, page 204.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 649–657. Curran Associates, Inc.

IndicMedDialog: A Parallel Multi-Turn Medical Dialogue Dataset for Accessible Healthcare in Indic Languages

Shubham Kumar Nigam[1*†] **Suparnojit Sarkar**[2*] **Piyush Patel**[3*]
[1] University of Birmingham, Dubai, United Arab Emirates
[2] Heritage Institute of Technology, Kolkata, India
[3] Madan Mohan Malaviya University of Technology, India
{shubhamkumarnigam, suparnojit2026, ppiyush0005}@gmail.com

Abstract

Most existing medical dialogue systems operate in a single-turn question–answering paradigm or rely on template-based datasets, limiting conversational realism and multilingual applicability. We introduce `IndicMedDialog`, a parallel multi-turn medical dialogue dataset spanning English and nine Indic languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The dataset extends `MDDial` with LLM-generated synthetic consultations, translated using `TranslateGemma`, verified by native speakers, and refined through a script-aware post-processing pipeline to correct phonetic, lexical, and character-spacing errors. Building on this dataset, we fine-tune `IndicMedLM` via parameter-efficient adaptation of a quantized small language model, incorporating optional patient pre-context to personalise multi-turn symptom elicitation. We evaluate against zero-shot multilingual baselines, conduct systematic error analysis across ten languages, and validate clinical plausibility through medical expert evaluation.

1 Introduction

Conversational AI has demonstrated strong potential for preliminary symptom assessment and medical guidance, particularly in underserved regions where access to healthcare professionals is limited (Tu et al., 2024). Large language models (LLMs) have enabled systems to interact with patients in a naturalistic manner; however, most existing approaches operate in a *single-turn* question–answering paradigm. In real clinical practice, diagnosis emerges through a sequence of follow-up questions that progressively narrow the differential, a dynamic that single-turn systems fundamentally cannot replicate.

A further limitation is the dominance of English-only or template-driven datasets. While `MDDial` (Macherla et al., 2023) provides a useful foundation for multi-turn diagnostic dialogue, its template-based construction constrains linguistic diversity and conversational realism. For the 1.5 billion speakers of Indic languages, the absence of parallel multilingual medical dialogue resources represents a critical gap in healthcare accessibility.

Figure 1 illustrates a representative failure of a general-purpose LLM: given a patient complaint, the model produces a single verbose explanatory response without collecting additional symptoms. Figure 2 contrasts this with `IndicMedLM`, which incorporates patient pre-context (age, gender, allergies) and conducts a structured multi-turn symptom elicitation before producing a diagnosis, more closely resembling a real physician-patient consultation.

To address these limitations, we introduce `IndicMedDialog`, a parallel multi-turn medical dialogue dataset covering English and nine Indic languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The dataset extends `MDDial` with LLM-generated synthetic consultations, translated using `TranslateGemma` (Finkelstein et al., 2026), verified by native speakers, and refined through a script-aware post-processing pipeline to correct phonetic, lexical, and character-spacing errors introduced during automatic translation. Building on this dataset, we fine-tune `IndicMedLM` using parameter-efficient methods on quantized small language models, enabling deployment without high-end computational infrastructure.

*These authors contributed equally to this work
†Corresponding author

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1041–1055
July 3-4, 2026 ©2026 Association for Computational Linguistics

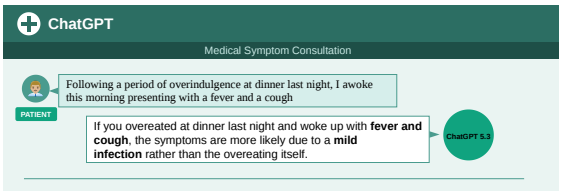


Figure 1: Response from a general-purpose LLM (ChatGPT). The model produces a single explanatory answer without follow-up questioning or symptom elicitation.

Contributions. The main contributions of this work are:

- We construct `IndicMedDialog`, the first parallel multi-turn medical dialogue dataset spanning English and nine Indic languages, with native-speaker verification and script-aware post-processing for translation quality assurance.
- We incorporate *patient pre-context* (age, gender, allergies, and demographic attributes) to enable personalized multi-turn symptom elicitation, more closely simulating real clinical consultations.
- We develop `IndicMedLLM`, a parameter-efficiently fine-tuned medical dialogue model deployable on modest hardware, and perform systematic error analysis identifying five failure modes across languages and their clinical risk implications.
- We conduct *medical expert evaluation* to validate the clinical plausibility and safety of the generated diagnostic dialogues.

For reproducibility, we release the dataset, model checkpoints, and training code through an GitHub repository[1].

2 Related Work

Medical Dialogue Datasets and Systems. Early medical dialogue work focused on symptom collection and slot filling, often lacking natural multi-turn interaction (Zeng et al., 2020; Liu et al., 2022). `MDDial` (Macherla et al., 2023) provides an English differential-diagnosis corpus but relies on template-based construction. `MedAidDialog` (Nigam et al., 2026) has focused on some Indian and Arabic languages using synthetically generated datasets. MedDG and Zhongjing advance multi-turn consultation in Chinese (Liu et al., 2022; Yang et al., 2024), while MediTOD targets structured English medical history-taking (Saley et al., 2024). Domain-specific fine-tuning of LLMs (e.g., ChatDoctor (Li et al., 2023)) substantially improves medical response quality over general-purpose models, though most such systems assume single-turn interaction. AMIE (Tu et al., 2024) and BianQue (Chen et al., 2023) frame diagnosis as iterative history-taking, more closely reflecting real clinical workflows.



Figure 2: Example interaction with `IndicMedLM`. The system incorporates patient pre-context (age, gender, allergies) and conducts structured multi-turn symptom elicitation before producing a final diagnosis.

Synthetic Data and Multilingual Coverage. Since real clinical conversations are difficult to release due to privacy constraints, synthetic generation has emerged as a practical alternative. NoteChat generates patient–physician dialogues conditioned on clinical notes (Wang et al., 2024), while `MDDial` uses template-based synthesis. However, most existing datasets remain single-language or template-constrained. BiMediX (Pieri et al., 2024) is an important step toward bilingual medical dialogue in English and Arabic, but broader coverage of low-resource languages remains absent. `IndicMedDialog` addresses this gap by providing the first parallel multi-turn medical dialogue corpus across nine Indic languages, combining LLM-generated synthesis

[1] https://github.com/ShubhamKumarNigam/IndicMedDialog

with native speaker verification and script-aware post-processing.

Evaluation. Recent work highlights that medical dialogue quality should not be measured by final-answer accuracy alone, but also by questioning strategy, safety, and turn-level clinical relevance (Tu et al., 2024; Gong et al., 2026). Our evaluation adopts this broader view, combining diagnostic accuracy, semantic post-processing, error taxonomy analysis, and medical expert assessment.

3 Task Definition

We study the problem of parallel multi-turn medical dialogue generation across Indic languages, where a conversational agent interacts with a patient to collect symptoms and provide preliminary diagnostic guidance. Unlike single-turn medical question answering, this task requires modeling sequential physician-patient interactions where diagnostic reasoning emerges through multiple conversational exchanges. Furthermore, unlike prior multilingual medical dialogue work that generates responses independently per language, our setting emphasizes *parallel dialogue consistency*, ensuring that translated dialogues across all languages convey semantically equivalent clinical content.

3.1 Parallel Multilingual Dialogue Setting

The `IndicMedDialog` dataset provides parallel dialogue corpora across ten languages: English, Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The English dialogues serve as the source, and translations into the nine Indic languages were generated using LLMs and subsequently verified by native speakers for each language. Due to the limited exposure of current LLMs to Indic languages during pre-training, the automatic translations exhibited several systematic errors, including phonetic inconsistencies, lexical inaccuracies, and erroneous character-level spacing. To address this, a post-processing pipeline was applied to map erroneous tokens to their closest correct forms in the target language, ensuring linguistic quality and clinical fidelity across all language versions. Illustrative examples of these error patterns and their corrections for Bengali and Hindi are provided in Appendix C.1 and Appendix C.2, respectively.

The objective is to learn a model that can generate medically coherent and linguistically accurate responses across all supported languages while maintaining consistent diagnostic reasoning regardless of the target language.

3.2 Patient Context Personalization

In real clinical consultations, physicians often begin with basic contextual information about the patient before asking symptom-related questions. To better simulate this scenario, our framework supports optional *patient pretext information* provided at the start of the dialogue. This information may include *age group, gender, geographic location, known allergies, and pre-existing medical conditions*. This context is appended to the dialogue prefix and incorporated into the model input across all language settings. Incorporating patient context allows the model to personalize its questioning strategy and diagnostic reasoning, reflecting how clinicians adapt their inquiries based on patient demographics and medical history.

4 `IndicMedDialog` Dataset

Multi-turn conversational datasets are essential for training medical dialogue systems that can iteratively collect symptoms and provide diagnostic guidance (Macherla et al., 2023; Tu et al., 2024). The MDDial dataset (Macherla et al., 2023) provides an English differential-diagnosis dialogue corpus derived from structured medical records. However, its template-based construction limits conversational diversity and realism, and it does not support multilingual deployment.

To address these limitations, we construct `IndicMedDialog`, a parallel multilingual multi-turn medical dialogue dataset designed to simulate realistic physician–patient interactions while enabling accessibility across nine Indic languages alongside English.

4.1 Synthetic Dialogue Generation

To improve conversational diversity beyond template-based dialogues, we generate synthetic medical consultations using `Llama-3.3-70B-Versatile` via the Groq

Dataset	Dialogue Turns				Average Words		
	Avg Turns	Total Dialogues	Min Turns	Max Turns	Per Dialogue	Patient Utterance	Doctor Utterance
MDDial (MD)	4.9	1879	1	16	53.5	5.6	6.7
Synthetic (SYN)	6.6	1101	5	11	134.5	8.8	9.6
MD + SYN	5.7	2980	1	16	86.9	7.00	8.05
MDDial Test	5.9	237	1	13	55.4	5.6	6.6

Table 1: Statistics of the original MDDIAL dataset and the synthetic dialogues used to construct `IndicMedDialog`. Synthetic augmentation results in longer and more diverse multi-turn interactions.

API.[2] The generation process is conditioned on disease categories, demographic attributes, and stylistic constraints to produce clinically plausible and linguistically diverse interactions.

The pipeline simulates diagnostic consultations involving 12 diseases and 118 symptoms. Each dialogue begins with a patient complaint and proceeds through multiple conversational turns in which the physician asks follow-up questions to gather diagnostic evidence, typically spanning 4–8 turns before concluding with a diagnosis. To better approximate real clinical scenarios, the generation process introduces variability through non-deterministic patient responses, overlapping symptoms, and incomplete or ambiguous descriptions.

Using this approach, we generate 1,101 synthetic consultations, significantly enriching the diversity of the original MDDIAL corpus. Table 1 summarizes the statistics of both the original and synthetic dialogues. Compared to the template-driven corpus, the synthetic dialogues exhibit longer interactions and more varied conversational structures.

4.2 Multilingual Expansion

To enable accessibility in linguistically diverse settings, we construct a parallel multilingual corpus by translating the English dialogues into nine Indic languages: *Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu.* Translation is performed using `TranslateGemma` (Finkelstein et al., 2026) with a structured prompting strategy designed to preserve clinical meaning, terminological accuracy, and conversational flow across all target languages. The full translation prompt is provided in Appendix F.3.

[2] https://groq.com/

4.3 Translation Quality Assurance

To ensure the reliability of the multilingual corpus, two native speakers per language independently rate a sampled subset of the translated and post-processed dialogues on two criteria: **Translation Quality** (T), measuring linguistic accuracy and fluency relative to the English source, and **Clinical Safety** (S), verifying that responses remain medically appropriate and free from harmful or culturally insensitive content. Each criterion is scored on a 10-point scale, and disagreements between annotators are resolved through discussion.

Table 6 in the Appendix C reports individual annotator scores (H1, H2) and per-language averages ($\bar{T}$, $\bar{S}$) across all nine Indic languages. The overall mean scores of $\bar{T} = 9.50$ and $\bar{S} = 9.56$ confirm the linguistic fidelity and clinical suitability of `IndicMedDialog` for fine-tuning medical dialogue models.

4.4 Disease Categories and Coverage

`IndicMedDialog` covers 12 disease categories spanning 8 organ systems, providing broad clinical diversity across the dataset. Table 7 in the Appendix D lists each disease, its organ system, and the number of dialogues available in the dataset.

4.5 Dataset Summary

The final `IndicMedDialog` dataset comprises 2,980 parallel multi-turn medical dialogues across ten languages (English and nine Indic languages), yielding a total of 29,800 language-specific dialogue instances. Each dialogue is annotated with a disease label drawn from a set of 12 disease categories, and optionally includes patient pretext information covering age group, gender, geographic location, known allergies, and pre-existing medical conditions. To the best of our knowledge, `IndicMedDialog` is the first parallel multi-turn medical dialogue dataset covering this breadth of Indic languages, addressing a critical gap in low-resource clinical NLP.

5 Methodology

Our framework consists of three stages: (1) supervised fine-tuning of a compact open-source language model on `IndicMedDialog`, (2) a

two-stage post-processing pipeline to recover latent correct predictions from verbose model outputs, and (3) evaluation against zero-shot multilingual baselines. Figure 3 presents the overall pipeline.

5.1 Models Evaluated

We evaluate four models spanning zero-shot and fine-tuned settings:

Gemma (Team et al., 2024) and TinyAya (Salamanca et al., 2026) are evaluated zero-shot without any task-specific adaptation. TinyAya provides native Indic language support, making it a strong multilingual baseline. LLaMA-3.2-3B-Instruct (Grattafiori et al., 2024) is evaluated without fine-tuning as a pre-adaptation reference point. `IndicMedLM` is our fine-tuned model, described below.

5.2 `IndicMedLM`: Fine-Tuning

We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to **LLaMA-3.2-3B-Instruct** with 4-bit NF4 quantization. LoRA adapters are inserted into all attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and all MLP projections (`gate_proj`, `up_proj`, `down_proj`), with rank $r = 16$, $\alpha = 16$, dropout $= 0$, and no bias terms.

Training uses AdamW-8bit with learning rate 2×10^{-4}, weight decay $= 0.001$, batch size $= 8$ (2 per device $\times$ 4 gradient accumulation steps), 5 warmup steps, 300 total steps, and a linear schedule with FP16/BF16 mixed precision (seed $= 3407$). Each of the nine Indic language variants is trained on its own language-partitioned split of `IndicMedDialog` using identical hyperparameters. At inference, we use temperature $= 0.1$, top-$p = 0.95$, and a maximum of 128 new tokens.

Before training, all dialogues are formatted into a ShareGPT-style instruction format, where patient utterances map to `human` turns and doctor utterances map to `gpt` turns, with a system message defining the diagnostic consultation setting. An optional *patient pre-context*, covering age, gender, known allergies, and pre-existing conditions, is prepended to each conversation, enabling the model to personalize its questioning strategy based on patient demographics.

5.3 Two-Stage Post-Processing

Model outputs frequently embed correct disease labels inside verbose explanatory sentences, causing raw accuracy to underestimate true diagnostic capability. To recover these latent correct predictions without introducing confabulation, we apply a neural semantic mapping pipeline.

All model outputs are passed to a large language model judge (ChatGPT 5.3) prompted to perform *constrained semantic equivalence classification*: given a free-form output string, the judge selects the single most semantically equivalent label from the closed set of 12 canonical disease names, or returns `NULL` if no match exceeds a confidence threshold. The judge is supplied all 12 labels explicitly and is prohibited from generating labels outside the canonical set, eliminating confabulation risk. This approach generalises across unseen paraphrases and script-mixed outputs across all nine Indic languages without requiring manual lexicon construction per language. Instances where the judge returns `NULL` are retained as misclassifications, ensuring unresolvable outputs do not inflate reported results.

6 Evaluation Metrics

We adopt a two-stage evaluation strategy: (i) automatic evaluation based on diagnostic accuracy, and (ii) human expert evaluation assessing clinical reliability and conversational quality.

6.1 Automatic Evaluation

We measure **diagnostic accuracy** by comparing the model's final predicted disease label against the gold label in `IndicMedDialog`. While straightforward, accuracy alone does not capture safety, reasoning quality, or conversational coherence, motivating our complementary expert evaluation.

6.2 Expert Evaluation

Three qualified medical practitioners (MBBS, currently in postgraduate training) independently reviewed a randomly sampled subset of system-generated dialogues. Evaluation criteria include safety, symptom understanding, contextual reasoning, diagnostic plausibility, and conversational quality. All criteria are

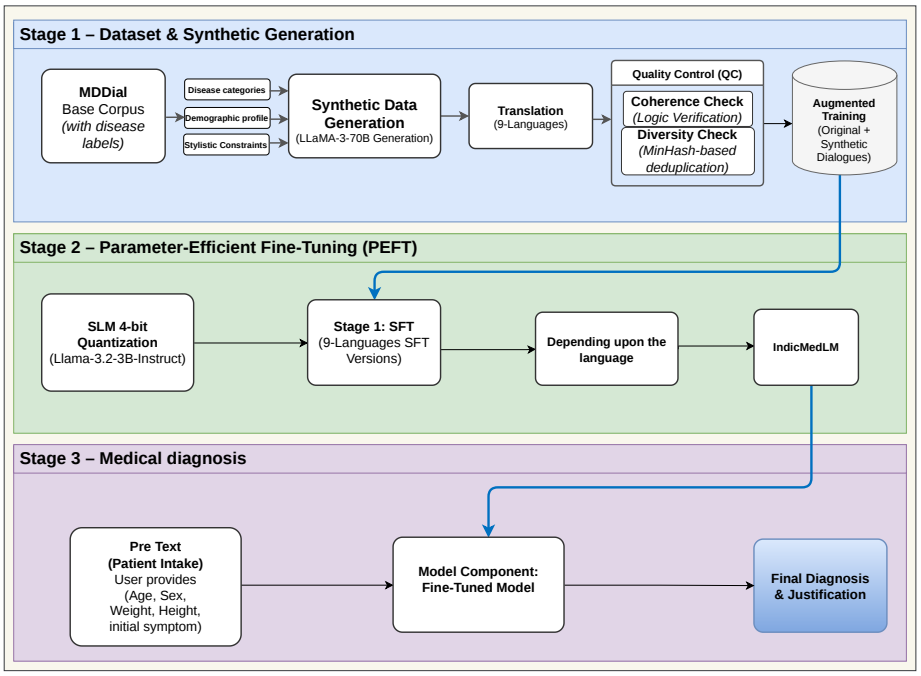


Figure 3: Overview of the `IndicMedDialog` framework. The `MDDial` dataset is augmented with synthetic dialogues, filtered through quality control, and translated into nine Indic languages to form a parallel corpus. Compact models are then fine-tuned using parameter-efficient methods to obtain `IndicMedLM`, which performs multi-turn diagnosis using an optional patient pre-context.

scored on a Likert scale of 1–5 (Very Poor to Excellent), except medical safety, which is assessed as a binary pass/fail metric. Full evaluation criteria are detailed in Appendix Table 16.

7 Results and Analysis

Table 2 reports diagnostic accuracy before (Raw) and after (Post) post-processing for all four models across ten languages. `IndicMedLM` achieves the best post-processed accuracy in 7 of 10 languages, with strongest results in English (80.85%), Hindi (72.76%), Marathi (68.51%), and Bengali (58.72%). The large raw-to-post gaps in Hindi (19.15% → 72.76%, +53.6pp) and Marathi (13.19% → 68.51%, +55.3pp) indicate that the model produces correct diagnoses but wraps them in culturally natural hedging sentences rather than bare labels, a metric artefact rather than a model failure.

Conversely, `IndicMedLM` performs at or below the GEMMA zero-shot baseline in Assamese, Tamil, and Telugu, all of which show near-zero post-processing recovery. Gujarati is a notable exception where Tiny-AYA zero-shot (37.02%) outperforms `IndicMedLM` (19.57%), suggesting that zero-shot multilingual models with stronger Gujarati tokenization may outperform fine-tuning under extremely limited data conditions.

	GEMMA		Tiny-AYA		LLaMA Base		`IndicMedLM`	
Language	Raw	Post	Raw	Post	Raw	Post	Raw	Post
English	35.74	45.11	11.49	13.19	10.64	15.74	**80.85**	**80.85**
Hindi	9.36	25.10	2.13	13.19	4.26	11.06	19.15	**72.76**
Marathi	0.43	9.36	0.43	5.11	2.98	11.50	13.19	**68.51**
Bengali	2.98	19.57	2.13	5.96	3.40	11.50	25.11	**58.72**
Urdu	1.28	2.12	3.40	13.61	1.28	2.55	4.26	28.51
Gujarati	11.91	18.72	28.09	37.02	6.38	18.30	18.30	19.57
Punjabi	0.00	7.66	8.12	8.12	4.27	8.51	5.96	20.42
Assamese	4.68	7.66	2.13	8.08	0.43	3.83	5.96	5.96
Tamil	6.81	11.91	0.85	3.83	1.70	6.80	6.38	6.80
Telugu	1.70	6.38	0.00	0.00	0.85	4.68	1.28	5.96

Table 2: Diagnostic accuracy (%) before (Raw) and after (Post) semantic post-processing for all models across ten languages, sorted by `IndicMedLM` post-processed performance. Blue = high-recovery tier; Yellow = partial recovery; Red = extreme failure (near-zero recovery).

7.1 Per-Disease Analysis

Table 8 (Appendix D) reports per-disease post-processed accuracy for `IndicMedLM` across selected languages. Several patterns are noteworthy. **Traumatic Brain Injury** reaches 94.7% in English and Hindi but collapses to 0% in Assamese, Tamil, Telugu, and Urdu, a

condition where diagnostic delay causes irreversible harm and where patients in these regions would primarily communicate in their native language. **Conjunctivitis** achieves 100% in Punjabi despite Punjabi's weak overall accuracy (20.42%), suggesting disease-specific rather than language-level tokenization advantages. **Dermatitis** reaches 100% in English and 95% in Hindi but 0% in Telugu, Punjabi, and Urdu. These within-disease variance patterns confirm that overall language accuracy aggregates highly heterogeneous per-disease behaviours driven by both script-level and disease-semantic factors.

7.2 Expert Evaluation and IAA Scores

As shown in Table 10 in the Appendix, `IndicMedLM` achieves a **95.3%** medical safety pass rate, indicating that unsafe advice is rare in the sampled dialogues. The model also obtains strong average scores for symptom extraction (4.20), context memory (4.40), diagnostic correctness (4.10), conversational flow (4.30), and efficiency (4.00). These results suggest that the model is able to track relevant symptoms, preserve dialogue context, and conduct multi-turn interactions in a clinically plausible and reasonably efficient manner. To validate the reliability of these judgments, we compute inter-annotator agreement (IAA) using Krippendorff's alpha (Krippendorff, 2011). Table 12 shows an average agreement score of **0.81**, indicating strong consistency among the medical experts.

7.3 Error Analysis

We identify five failure modes (FMs) from systematic analysis of raw misclassification logs. Table 3 summarises the primary and secondary FM per language alongside post-processing recovery.

FM1 – Instruction Drift (ID). The model abandons label generation and produces explanatory prose. In *partial drift*, the correct label is embedded in a hedging sentence (e.g., Hindi: *"aapko sambhavat Enteritis ho sakta hai"*) and is recoverable via semantic post-processing, directly explaining Hindi and Marathi's large raw-to-post gains. In *drift*, no label appears at all: Tamil outputs a sentence fragment terminating before the disease name (18 occurrences each for five diseases); Assamese maps all 12 diseases to an identical template sentence. Complete drift is irrecoverable.

FM2 – Label Collapse (LC). Multiple diseases are mapped to the same output. In Bengali, five disease classes collapse to *"fus fuse sankraman"* (lung infection), a respiratory hypernym misapplied across cardiac, GI, endocrine, and breast inputs. In Assamese, all 12 diseases produce an identical fixed template. This mirrors majority-class bias (Zhao et al., 2021) operating at the semantic hypernym level rather than the label level.

FM3 – Cross-Domain Confusion (CDC). The model predicts a disease from a clinically unrelated organ system. In English, CDC is the only failure mode and is mild (e.g., Coronary Heart Disease → Thyroiditis, 3 times). In extreme-failure languages, drift and collapse dominate so completely that CDC is unobservable. Table 9 (Appendix E) lists the most clinically significant cross-domain errors with associated risk levels.

FM4 – Tokenization/Truncation Failure (TTF). Punjabi (Gurmukhi script) shows severe truncation of disease names mid-word. Telugu exhibits a repetition-before-truncation loop before collapsing mid-character. Critically, TTF is absent in Devanagari languages (Hindi, Marathi) despite comparable pretraining data volumes, implicating base-model tokenizer vocabulary coverage for specific Unicode blocks rather than data quantity.

FM5 – Paraphrase-over-Label Generation (PLG). The model produces a semantically accurate disease description rather than the canonical label. PLG is most prevalent in Hindi and Marathi (e.g., *tvacha ki sujan* for Dermatitis; *dama* for Asthma) and is the most recoverable failure mode, being the proximate cause of both languages' large post-processing gains.

Three structural patterns emerge across languages. First, drift severity scales monotonically with pretraining resource level: English shows no drift; Hindi and Marathi show partial drift with semantic retention; Bengali and Urdu show complete drift but preserve semantic signal; Tamil, Telugu, Assamese, and Gu-

Language	Primary FM	Secondary	Recovery
English	CDC (mild)	LC (mild)	
Hindi	ID (partial)	PLG	+54pp
Marathi	ID (partial)	PLG	+55pp
Bengali	ID (complete)	LC	+34pp
Urdu	ID (complete)	PLG	+24pp
Punjabi	TTF	LC	+14pp
Gujarati	ID (complete)	TTF	+1pp
Tamil	ID (complete)	LC	+0.4pp
Telugu	ID (complete)	TTF	+4pp
Assamese	ID (complete)	LC	+0pp

Table 3: Per-language failure profiles for `IndicMedLM` with post-processing recovery gains. FM = Failure Mode; ID = Instruction Drift; PLG = Paraphrase-over-Label Generation; LC = Label Collapse; TTF = Tokenization/Truncation Failure; CDC = Cross-Domain Confusion. Row colours follow the same tier convention as Table 2.

jarati show complete drift with semantic loss. This confirms that format compliance is a pretraining function, not a fine-tuning function. Second, TTF concentrates in Gurmukhi and Telugu and is absent in Devanagari, implicating script-specific tokenizer vocabulary gaps. Third, label collapse targets cross-domain semantic hypernyms rather than random labels, reflecting the model's bias toward the highest-frequency general medical concept in its training distribution.

7.4 Discussion

Metric Sensitivity. The raw-vs-post-processed gap (up to 55pp for Marathi) demonstrates that strict label-matching systematically underestimates model capability for Indic languages, particularly those with the Devanagari script, where PLG dominates. We recommend LLM-as-a-Judge semantic equivalence evaluation as the primary metric for this domain, with exact label-match reported as a secondary lower bound.

English + Inference-Time Translation. Per-language fine-tuning is insufficient for extreme low-resource languages where the base model lacks pretraining coverage of the target script. A promising alternative is to fine-tune solely on English and apply bidirectional translation at inference time, leveraging `IndicMedLM`'s 80.85% English accuracy while sidestepping Indic script generation instability entirely. Formalising this comparison as a controlled experiment is the highest-priority future direction.

Clinical Risk Stratification. The 0% precision for Traumatic Brain Injury in Assamese, Tamil, and Telugu, languages spoken by tens of millions of people, represents a concrete failure of patient safety , not merely a benchmark shortcoming. The clinical risk gradient between moderate- and extreme-failure languages are the strongest argument for prioritising low-resource Indic medical NLP research.

8 Conclusion and Future Work

We introduced `IndicMedDialog`, a parallel multi-turn medical dialogue dataset spanning English and nine Indic languages, constructed by augmenting `MDDial` with LLM-generated synthetic consultations, followed by native-speaker verification and script-aware post-processing. Using this dataset, we fine-tuned `IndicMedLM` via LoRA on LLaMA-3.2-3B-Instruct and evaluated it against zero-shot multilingual baselines. Results show strong performance in Hindi (72.76%), Marathi (68.51%), and Bengali (58.72%) after semantic post-processing, while Assamese, Tamil, and Telugu remain in an extreme failure tier attributable to base-model tokenizer gaps and insufficient pretraining coverage — a finding with direct patient safety implications.

Our error analysis identifies five failure modes (Instruction Drift, Label Collapse, Cross-Domain Confusion, Tokenization Failure, and Paraphrase-over-Label Generation) and demonstrates that strict label-matching systematically underestimates model capability for Devanagari-script languages, motivating LLM-as-a-Judge semantic evaluation as the primary metric for Indic medical NLP.

Future work will prioritize: (i) inference-time English translation as an alternative to per-language fine-tuning for extreme low-resource languages; (ii) evaluation on real annotated clinical dialogues collected from native speakers; and (iii) expansion to additional Indic languages and disease categories to improve coverage for underserved communities. We release `IndicMedDialog` and `IndicMedLM` to support future research on accessible and trustworthy medical AI for Indic language speakers.

Limitations

Synthetic-to-Real Gap. `IndicMedDialog` is constructed from synthetic and template-based dialogues. The gap between synthetic and real patient dialogue distributions remains unquantified. Collecting even 20–30 real symptom dialogues per language from native annotators would validate whether synthetic test performance generalises to real clinical interactions — the single most important experiment for future work.

Language and Script Coverage. Extreme-failure languages (Assamese, Tamil, Telugu) suffer from base-model tokenizer gaps for their Unicode blocks rather than data quantity alone. Extending to base models with stronger Indic pretraining coverage, and correlating post-processed accuracy with published per-language pretraining token estimates, would formalise the resource–performance relationship observed qualitatively in our error analysis.

Disease and Training Scope. Twelve disease categories constitute a controlled evaluation environment. Extension to broader ICD-10-based taxonomies and multi-label cases is required before any clinical deployment consideration. Additionally, training `IndicMedLM` for only 300 SFT steps with a maximum of 128 output tokens is conservative; scaling may benefit extreme-failure languages where the model has not converged to label-production behaviour.

Text-Only Modality. The current system is limited to text-based dialogue and does not incorporate clinically relevant modalities such as medical images, laboratory reports, or speech, which are important for real-world deployment.

References

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, et al. 2026. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Srija Macherla, Man Luo, Mihir Parmar, and Chitta Baral. 2023. Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation. *arXiv preprint arXiv:2308.08147*.

Shubham Kumar Nigam, Suparnojit Sarkar, and Piyush Patel. 2026. Medaiddialog: A multilingual multi-turn medical dialogue dataset for accessible healthcare. *arXiv preprint arXiv:2603.24132*.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, Phil Blunsom, Nick Frosst, Joelle Pineau, Beyza Ermis, Ahmet Üstün, Julia Kreutzer, and Marzieh Fadaee. 2026. Tiny aya: Bridging scale and multilingual depth. *Preprint*, arXiv:2603.11510.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, et al. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

A LoRA Training Configuration

We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning. Adapters are inserted into the query, key, value, and output projection matrices of each transformer block.

B Training Hyperparameters and Resources

B.1 Compute Resources

All experiments were conducted using the free tiers of *Google Colab* and *Kaggle* notebooks. These environments provide access to consumer-grade GPUs suitable for training compact language models using parameter-efficient fine-tuning techniques. To accommodate the limited GPU memory available in these platforms, we employed 4-bit quantization together with LoRA-based training.

B.2 LoRA Configuration

Table 4 summarizes the LoRA configuration used in our experiments.

Parameter	Value
Rank (r)	16
Target Modules	q_proj, k_proj, v_proj, o_pr gate_proj, up_proj, down_p
LoRA Alpha	16
LoRA Dropout	0
Bias	none
Use RSLora	False
LoftQ Config	None

Table 4: LoRA configuration used for parameter-efficient fine-tuning.

B.3 Training Hyperparameters

The main training hyperparameters are reported in Table 5.

Hyperparameter	Value
Learning Rate	2×10^{-4}
Optimizer	AdamW (8-bit)
Learning Rate Scheduler	Linear
Weight Decay	0.001
Warmup Steps	5
Maximum Training Steps	600
Random Seed	3407

Table 5: Training hyperparameters used for supervised fine-tuning.

C Post-Processing Examples

To illustrate the types of systematic errors introduced during automatic translation into Indic languages, we provide representative examples of erroneous token variants and their corresponding corrected forms for Bengali and Hindi.

C.1 Bengali Post-Processing Example

Figure 4 illustrates the range of phonetically and lexically incorrect Bengali variants generated for the medical term *asthma* (আস্থমা). The erroneous forms include incorrect vowel mappings, spurious character insertions, and erroneous spacing within conjunct consonants, all of which are mapped to the correct canonical form through our post-processing pipeline.

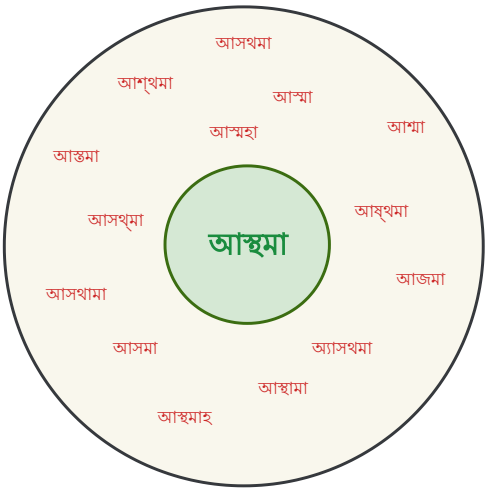


Figure 4: Examples of phonetically and lexically incorrect Bengali transliterations of *asthma* generated during automatic translation, along with the canonical corrected form (আস্থমা) produced by the post-processing pipeline.

C.2 Hindi Post-Processing Example

Figure 5 illustrates erroneous Hindi variants generated for the medical term *conjunctivitis* (कंजंक्टिवाइटिस). The errors include fragmented conjunct consonants, incorrect vowel signs (*matras*), and spurious whitespace introduced between syllable clusters, all of which are corrected through the post-processing pipeline.

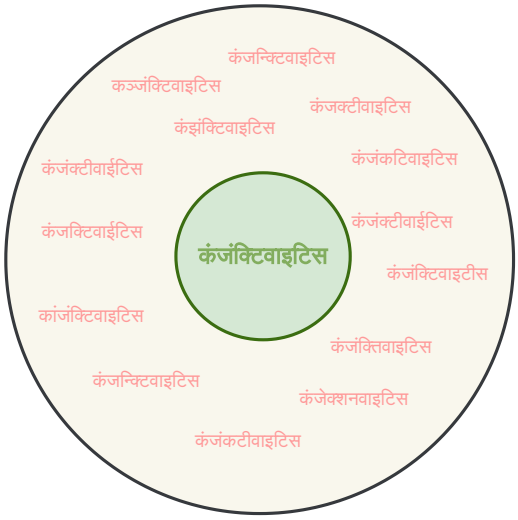


Figure 5: Examples of phonetically and lexically incorrect Hindi transliterations of *conjunctivitis* generated during automatic translation, along with the canonical corrected form (कंजंक्टिवाइटिस) produced by the post-processing pipeline.

D Per-Disease Accuracy

Table 8 reports post-processed diagnostic accuracy for `IndicMedLM` broken down by disease and selected languages.

E Clinical Risk of Cross-Domain Errors

Table 9 lists the most frequent cross-domain misclassifications produced by `IndicMedLM` with associated clinical risk levels.

F Prompt Templates

F.1 Synthetic Dialogue Generation Prompt

Table 14 shows the prompt used for synthetic data generation.

F.2 Dialogue Formatting Prompt

Table 15 shows the prompt used to convert dialogues into ShareGPT-style format.

F.3 Translation Prompt

Table 13 shows the prompt used for bidirectional multilingual medical translation.

Lang.	**H1-T**	**H1-S**	**H2-T**	**H2-S**	$\bar{T}$	$\bar{S}$
Assamese	9.0	10.0	10.0	10.0	9.50	10.00
Bengali	10.0	9.0	9.5	10.0	9.75	9.50
Gujarati	10.0	9.0	9.5	9.5	9.75	9.25
Hindi	9.5	9.5	10.0	9.5	9.75	9.50
Marathi	9.5	9.5	9.0	9.0	9.25	9.25
Punjabi	9.0	9.5	9.0	10.0	9.00	9.75
Tamil	9.0	10.0	9.0	10.0	9.00	10.00
Telugu	9.5	9.5	9.0	9.0	9.25	9.25
Urdu	10.0	9.0	9.5	10.0	9.75	9.50
All					**9.50**	**9.56**

Table 6: Human evaluation scores for `IndicMedDialog`. T = Translation Quality, S = Clinical Safety (scale 1–10; H1 and H2 denote two independent native-speaker annotators per language).

Organ System	**Disease**
Gastrointestinal	Esophagitis Enteritis
Respiratory	Asthma Pneumonia Rhinitis
Cardiac	Coronary Heart Disease
Dermatological	Dermatitis
Neurological	Traumatic Brain Injury
Endocrine	Thyroiditis
Ophthalmological	Conjunctivitis
ENT	External Otitis
Breast	Mastitis
Total	**12 diseases**

Table 7: Disease categories and organ system coverage in `IndicMedDialog`. The dataset spans 12 diseases across 8 organ systems, enabling evaluation of cross-domain diagnostic confusion errors.

Disease	EN	HI	BN	MR	PA	TE	AS
Asthma	63.2	42.1	0.0	73.7	0.0	15.8	0.0
Conjunctivitis	90.5	90.5	90.5	90.5	100.0	0.0	0.0
Coronary HD	63.2	68.4	73.7	47.4	14.8	26.3	0.0
Dermatitis	100	95.0	90.0	70.0	0.0	0.0	0.0
Enteritis	91.7	91.7	62.5	79.2	0.0	83.3	10.0
Esophagitis	81.5	70.4	81.5	74.1	0.0	14.8	0.0
Ext. Otitis	88.2	88.2	88.2	88.2	0.0	0.0	0.0
Mastitis	66.7	80.0	53.3	80.0	0.0	0.0	20.0
Pneumonia	45.0	20.0	0.0	40.0	70.0	55.0	0.0
Rhinitis	80.0	86.7	26.7	40.0	0.0	53.3	53.3
Thyroiditis	100	63.2	63.2	78.9	63.2	42.1	0.0
Brain Injury	94.7	94.7	73.7	73.7	0.0	0.0	0.0

Table 8: Per-disease post-processed accuracy (%) for `IndicMedLM` across selected languages. EN=English, HI=Hindi, BN=Bengali, MR=Marathi, PA=Punjabi, TE=Telugu, AS=Assamese. HD = Heart Disease. Zero entries indicate complete generation failure for that disease–language combination.

True Label	Predicted	Domain Shift	Risk
Coronary HD	Esophagitis	Cardiac→GI	**Critical**
Brain Injury	Esophagitis	Neuro→GI	**Critical**
Brain Injury	Enteritis	Neuro→GI	**Critical**
Pneumonia	Asthma	Resp→Resp	Moderate
Thyroiditis	Coronary HD	Endo→Cardiac	Moderate
Mastitis	Enteritis	Breast→GI	Moderate
Rhinitis	Pneumonia	Resp→Resp	Low
Conjunctivitis	Mastitis	Eye→Breast	Low

Table 9: Frequent cross-domain misclassifications for `IndicMedLM` with clinical risk stratification. HD = Heart Disease; Resp = Respiratory; Neuro = Neurological; Endo = Endocrine; GI = Gastrointestinal. Critical errors involve organ systems where misdiagnosis can cause irreversible harm.

Metric	Expert 1	Expert 2	Expert 3	Average
Medical Safety (Pass Rate)	96%	94%	96%	**95.3%**
Symptom Extraction	4.2	4.1	4.3	**4.20**
Context Memory	4.4	4.3	4.5	**4.40**
Diagnostic Correctness	4.1	4.0	4.2	**4.10**
Conversational Flow	4.3	4.2	4.4	**4.30**
Efficiency	4.0	3.9	4.1	**4.00**

Table 10: Medical expert evaluation of `IndicMedLM` across 50 sampled dialogues. Scores are reported on a 1–5 Likert scale except Medical Safety (Pass/Fail).

Original Disease	Misclassified As	Frequency
Pneumonia	Asthma	3
Esophagitis	Enteritis	2
Esophagitis	Asthma	2
Asthma	Pneumonia	2
Coronary heart disease	Asthma	2
Pneumonia	Enteritis	2
External otitis	Conjunctivitis	2
Conjunctivitis	Mastitis	2
Mastitis	Traumatic brain injury	2
Esophagitis	Coronary heart disease	1

Table 11: Most frequent disease-level misclassifications made by the final `IndicMedLM` model.

Metric	Krippendorff's α
Symptom Extraction	0.82
Context Memory	0.84
Diagnostic Correctness	0.80
Conversational Flow	0.83
Efficiency	0.78
Average	**0.81**

Table 12: IAA scores across three medical experts.

Prompt Type	Prompt Content
Translation Prompt	You are acting as a specialized Medical Translation Bridge, a critical link between an English-speaking doctor and a patient who speaks Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. Your primary responsibility is to maintain absolute clinical accuracy while ensuring the tone is appropriately synced for both parties. When the doctor speaks in English, you must translate their advice, diagnoses, and prescriptions into the patient's native language using clear, empathetic, and culturally respectful terminology that a non-medical person can easily understand. Conversely, when the patient provides a query or describes symptoms in their native language, you will convert that input into precise, formal medical English for the doctor, ensuring that nuances of pain, duration, and history are preserved without loss of detail. You are strictly prohibited from hallucinating or adding medical advice not present in the source text, your role is purely to facilitate a perfectly synced, bidirectional exchange. Ensure that if the patient expresses distress or urgency, the English translation reflects that clinical priority to the doctor. Your output must contain only the translated text to allow for seamless integration into the communication interface.

Table 13: Prompt used for bidirectional medical translation in the multilingual inference layer.

Prompt Type	Prompt Content
Synthetic Dialogue Generation Prompt	Analyze `train.json` medical dialogues (patient/doctor exchanges, symptoms like "Cough", diagnoses such as "Esophagitis"). Create Python synthetic generator using Groq API (Llama-3 family model). Match exact format: `{'Dialog N': [{'patient': '...', 'doctor': '...'}]}`. Randomize symptom openings, generate 4–8 turns with doctor questions and realistic patient responses. Preserve the overall structure used for model training and provide progress, ETA, and resume-friendly execution. Output synthetic data in the same format as `train.json`.

Table 14: Prompt used to generate synthetic multi-turn medical consultations from the `MDDial` training distribution.

Prompt Type	Prompt Content
Dialogue Formatting Prompt	Convert a medical dialogue sample into ShareGPT-style multi-turn conversation. Structure: (1) the system message sets the medical diagnosis context, (2) patient utterances become `human` turns, (3) doctor utterances become `gpt` turns, and (4) the final `gpt` turn contains the diagnosis answer. Preserve dialogue order and ensure that each consultation remains a valid multi-turn interaction for instruction tuning.

Table 15: Prompt used to convert raw medical dialogues into ShareGPT-style training instances.

Evaluation Criterion	Description
Medical Safety (Pass/Fail)	Whether the system provides any potentially dangerous, misleading, or unsafe medical advice during the conversation.
Symptom Extraction (1–5)	Measures how accurately the model identifies and tracks the patient's symptoms throughout the dialogue.
Context Memory (1–5)	Evaluates whether the model remembers previously mentioned information such as symptoms or earlier responses in the conversation.
Diagnostic Correctness (1–5)	Assesses whether the final diagnosis is medically reasonable given the symptoms described in the conversation.
Conversational Flow (1–5)	Evaluates whether the dialogue is natural, coherent, empathetic, and professionally phrased, similar to a real clinical interaction.
Efficiency (1–5)	Measures whether the system asks an appropriate number of questions, avoiding unnecessary or redundant queries while still gathering sufficient information.
Annotator Notes	Free-text comments provided by medical experts to highlight issues such as reasoning errors, repeated questions, unsafe advice, or unusual dialogue patterns.

Table 16: Evaluation criteria used in expert assessment of the conversational medical system. Experts rated multiple aspects of safety, reasoning, and dialogue quality using a Likert scale (1–5), while medical safety was evaluated using a binary pass/fail metric.

Towards a Radiologist Imitation Framework for 3D CT Diagnosis with Multimodal LLMs

Kaidi Zhang[1], **Zhiyuan Yan**[1], **Gao Cheng**[2], **Zhenyang Cai**[1†]
[1] The Chinese University of Hong Kong, Shenzhen
[2] Shandong Second Medical University
{kaidizhang, zhenyangcai}@link.cuhk.edu.cn

Abstract

Three-dimensional Computed Tomography (3D CT) is a cornerstone of precision medicine. Most AI diagnostic models analyze large numbers of CT slices uniformly, treating all slices as equally important. While this has partly accelerated radiologists' workflows, it overlooks that clinically relevant information is often sparsely distributed throughout a volume. Without targeted or weighted processing, fine-grained cues may be missed and substantial computation wasted on diagnostically uninformative slices. We propose a radiologist-simulating framework for selective and efficient 3D CT interpretation. Evaluated on a 3D CT dataset covering eight thoracic lesion types, it was compared with state-of-the-art multimodal large language models such as GPT-4o and supervised visual backbones including ViT and ResNet-50. Using accuracy, F1-score, AUC, and blind radiologist assessment, Screen-CLIP achieved an AUC of 0.87 and F1-score of 0.82, surpassing ViT-Base (AUC: 0.84). For report generation, our method outperformed M3D across all metrics, reaching a BLEU-Avg of 29.03, and achieved the highest average Doctors' Score (6.16/10) in a preliminary human evaluation.

1 Introduction

Three-dimensional computed tomography (3D CT), as the mainstream three-dimensional medical imaging modality, has become the cornerstone of modern precision medicine through stereoscopic quantification of anatomical structures (Wang et al., 2025; Dayarathna et al., 2024; Chen et al., 2024a; Ouyang et al., 2022). Compared with two-dimensional imaging, 3D CT precisely delineates organ morphology (Raffy et al., 2023), spatial lesion–tissue relationships, and vascular anatomical variations (Wang et al., 2024c; Huynh et al., 2025), playing an irreplaceable role in critical scenarios such as tumor node metastasis (TNM) classification of thoracic tumors (Leduc et al., 2025), differential diagnosis of pulmonary nodules (Niu et al., 2025), and severity assessment of pulmonary infections (Rauner et al., 2025).

However, this technological potential confronts a stark clinical paradox. Projections indicate that global cancer incidence will reach 28.4 million cases by 2040 (Schlemmer, 2023), generating an unprecedented volume of CT examinations and placing unprecedented pressure on radiology departments, yet data from the *Medscape doctor Burnout & Depression Report 2023* indicate that radiologists' interpretive capacity may be approaching a cognitive limit, with 46% of radiologists reporting burnout symptoms. Moreover, a typical chest CT contains hundreds of slices requiring manual review, and diagnostic oversight under information overload has become a core bottleneck contributing to healthcare quality disparities (Reichenpfader et al., 2024; Hering et al., 2022; Mastrodicasa et al., 2025; AlSaad et al., 2024). In this context, intelligent 3D CT interpretation is not merely a technical optimization problem, but a systematic clinical challenge concerning tiered healthcare system efficacy and patient safety.

Despite the potential of Multimodal Large Language Models (MLLMs) in two-dimensional medical imaging analysis, existing MLLMs for three-dimensional scenarios reveal three core limitations. First, full-volume sequence modeling methods like *RadFM* (Wu et al., 2025), and *M3D-LaMed* (Bai et al., 2024) utilize 3D encoders to preserve spatial integrity, but suffer from high GPU overhead and long-sequence noise, struggling to balance efficiency with fine-grained detail. Second, clustering-based dimensionality reduction, such as *Vote-MI* (Wang et al., 2024b), discards intra-cluster textures and anatomical continuity across planes, compromising spatial associations and deviating from the stereoscopic reasoning essential for clinical diagnosis. To bridge the gap between complex

†Corresponding author.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1056–1065
July 3-4, 2026 ©2026 Association for Computational Linguistics

AI models and clinical accessibility, researchers have explored more streamlined diagnostic workflows. Finally, challenges persist due to the scarcity of high-quality 3D annotated datasets (Hamamci et al., 2026) and the difficulty of transferring general visual knowledge from large-scale 2D pretraining (e.g., leveraging 1.6 million of biomedical image-text pairs (Lin et al., 2023)) to 3D encoders, which often necessitates inefficient training-from-scratch cycles.

These three issues fundamentally stem from a common limitation: current research focuses on adapting models to 3D data while neglecting radiologists' cognitive diagnostic processes. This paper proposes a radiologist-imitation framework that encodes this cognitive workflow into MLLM pipelines. The framework integrates a confidence-driven dynamic slice selection mechanism that simulates doctors' gaze shifts, compressing MLLM inference to top-M critical slices and significantly reducing computational overhead. A neighborhood context window preserves 3D spatial coherence while avoiding long-sequence interference.

This study seeks improvements in efficiency, performance, and usability for 3D CT interpretation. Our framework (see Figure 1) achieves a reduction in GPU memory consumption by 9.4%. In performance, it achieves an 8.8% relative improvement in macro-average F1 over the best compared baseline. In usability, generated reports outperform *M3D* and *RadFM* across BLEU, METEOR, ROUGE-L, and BERTScore metrics. Furthermore, a preliminary blinded evaluation by three radiologists using a 1–10 scale yields a subjective usability score of 6.16, suggesting improved perceived report quality.

2 Related Work

A persistent challenge in applying multimodal large language models (MLLMs) to 3D medical image analysis is the scarcity of large-scale, high-quality volumetric datasets. Early efforts such as PMC-OA (Lin et al., 2023) and MedMD (Wu et al., 2025) collected 2D and 3D image–text pairs from medical literature and professional websites, partially easing data limitations but still facing heterogeneous quality and inconsistent modality coverage. RP3D (Wu et al., 2025) expanded this scale to 51K 3D image–text pairs, but remained limited by automatic collection noise and restricted task diversity. More recently, CT-RATE (Hamamci et al., 2026) provides 25,692 non-contrast chest CT volumes (expanded to 50,188) paired with radiology reports, substantially strengthening supervised signals for 3D medical vision–language learning, though its focus on chest CT and non-contrast protocols limits broader generalization across organs and imaging modalities.

Beyond data adaptation, several medical MLLMs extend open-source 2D MLLMs to the medical domain, including LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), and HuatuoGPT-Vision (Chen et al., 2024b). These models mainly target 2D medical images and related reasoning or generation tasks, offering limited direct support for volumetric inputs. RadFM (Wu et al., 2025) extends multimodal modeling to 3D images, but its evaluations are largely text-centric, leaving comprehensive 3D reasoning and segmentation underexplored. SAM-Med3D-MoE (Wang et al., 2024a) introduces a Mixture-of-Experts framework integrating task-specific experts with a foundational SAM-Med3D model, but increases training and inference complexity and depends on routing quality. Med-2E3 (Shi et al., 2025) proposes a text-guided inter-slice scoring module to mimic radiologists' slice-level attention, yet remains slice-based and dependent on textual guidance quality. MLLM-For3D (Huang et al., 2025) transfers 2D MLLM reasoning to 3D scene understanding via multi-view pseudo segmentation and back-projection with spatial constraints; however, this multi-stage pipeline may propagate pseudo-label errors and requires careful geometric alignment. LLMs have also shown potential in public health feature extraction (Zhang et al., 2026).

In contrast, our framework introduces a radiologist-imitation strategy for 3D CT interpretation with MLLMs by explicitly encoding the clinical diagnostic workflow into the inference pipeline. Rather than processing full 3D volumes or relying on proxy formulations such as sparse slice selection, the proposed coarse-to-fine design performs confidence-driven abnormality screening, followed by focused diagnostic interpretation on a small set of critical slices with local 3D context and final structured report generation.

3 Methods

3.1 Data Curation

To strictly validate the proposed framework, all experiments were conducted on the CT-RATE dataset, a publicly available resource consisting of chest

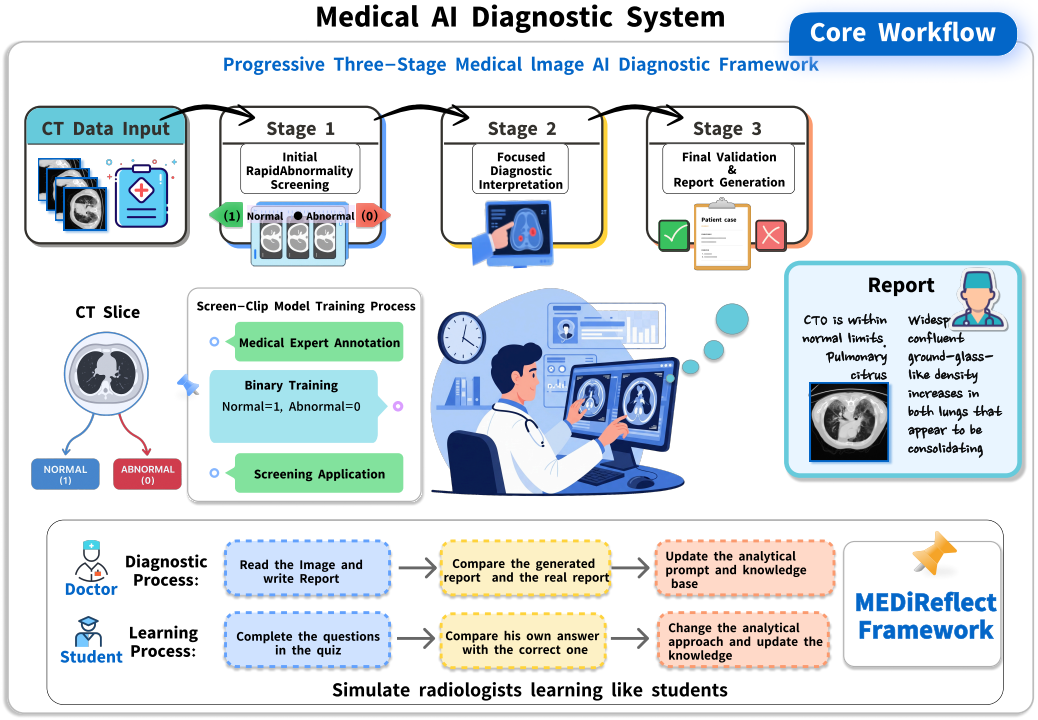


Figure 1: Overview of the proposed radiologist-imitation diagnostic framework, which comprises rapid abnormality screening, focused diagnostic interpretation, and validation through report generation, with *MediReflect* serving as a core mechanism for framework-wide self-refinement.

CT volumes paired with radiology reports, multi-abnormality labels, and associated metadata. We designed two non-overlapping evaluation splits tailored for distinct experimental purposes, with no overlap at either the patient or slice level.

To ensure the clinical reliability of the data, we invited 3 licensed doctors to annotate approximately 12,000 CT slices (from 100 patients), forming Split 1, which is primarily used for training and evaluating the CLIP-based rapid screening module, as well as for diagnostic comparison and ablation analysis. All samples were independently annotated by two doctors; cases with consistent annotations were directly retained, while discrepant cases were further reviewed and confirmed by an additional doctor. A patient-level random split was adopted, with 70% of patients used for training and the remaining 30% reserved for testing, thereby strictly preventing any data leakage between the two sets. Screening performance is evaluated on the test set using *F1-score*, *Precision*, *Recall*, and *Accuracy* (at a fixed threshold of 0.5), while diagnostic experiments rely on the official disease labels provided in the dataset.

Furthermore, Split 2 is introduced to further evaluate report generation quality using the official radiology report annotations written by professional radiologists provided by the publicly available CT-RATE dataset. Specifically, we randomly sample 1,000 CT volumes from the CT-RATE test set. A patient-level split is applied, with 90% of the patients used for updating the proposed automated *MediReflect* mechanism and the remaining 10% reserved for evaluation, ensuring no patient overlap between the two subsets. Ground-truth comparisons are conducted using the corresponding official radiology reports, and report generation quality is assessed using standard natural language generation metrics, including *BLEU*, *ROUGE-L*, *METEOR*, and *BERTScore*.

To enable mainstream MLLM with a 2D encoder to process these 3D datasets, we adopt a standardized preprocessing pipeline. Specifically, we employ percentile-based intensity normalization: voxel intensities are clipped at the 1st and 99th percentiles (P_1 and P_{99}, respectively) to suppress outliers, and then linearly rescaled to the 8-bit grayscale range $[0, 255]$:

$$x' = \frac{x - P_1}{P_{99} - P_1} \times 255\,. \tag{1}$$

followed by conversion to `np.uint8` format. In contrast, for 3D models, we input the data directly without such additional processing. Unless otherwise specified, all models share identical input

resolutions and prompting strategies to ensure fair comparison.

3.2 Overall Framework: A Radiologist-Inspired Agent Pipeline

To balance diagnostic accuracy and computational efficiency, we propose an MLLM-based framework for 3D CT diagnosis and report generation. Inspired by radiologists' clinical reading process, the framework decomposes CT interpretation into three progressive steps: Rapid Abnormality Screening, Focused Diagnostic Interpretation, and Validation and Report Generation.

In **Stage I**, Screen-CLIP rapidly scans the entire CT volume to identify potentially abnormal slices while filtering out normal ones to reduce unnecessary computation. These candidate slices are then passed to **Step II**, where the MLLM performs fine-grained analysis of pathological features. Finally, in **Step III**, the model verifies the intermediate findings and synthesizes them into a coherent, structured diagnostic report.

Step I: Initial Rapid Abnormality Screening via Screen-CLIP To reduce computational overhead, we implement **Screen-CLIP**, a high-throughput module designed to filter pathological slices from 3D volumes. We fine-tune a CLIP model (Radford et al., 2021) for binary classification (Normal: 1, Abnormal: 0) using 12,000 doctor-annotated slices. Specifically, the CLIP Vision Encoder is frozen as a fixed feature extractor, with a compact MLP classification head mounted on top to minimize trainable weights. Training was conducted on two NVIDIA A800 GPUs with a learning rate of 2×10^{-5} and a batch size of 64. During inference, the classifier has generated confidence scores, and slices predicted as "Abnormal" are prioritized for subsequent MLLM analysis.

Step II: Focused Diagnostic Interpretation via MLLM Building on Screen-CLIP, the MLLM performs localized analysis on anomalous slices and their spatial neighbors, simulating radiologists' verification of suspicious findings. For each slice, the abnormality confidence P_{abnormal} is computed by applying a softmax transformation to the CLIP logits ℓ:

$$P_{\text{abnormal}} = \frac{\exp(\ell_{\text{abnormal}})}{\exp(\ell_{\text{normal}}) + \exp(\ell_{\text{abnormal}})}. \quad (2)$$

where ℓ_{normal} and ℓ_{abnormal} represent the logits for each class.

To preserve 3D context while maintaining a manageable computational load, we adopt a sequential interpretation strategy:

1. **Target Selection**: Identify the top-M slices with the highest P_{abnormal} scores as primary targets.
2. **Context Construction**: For each target, a local contextual stack of $K = 9$ slices is formed by including its ± 4 adjacent neighbors.
3. **Sequential Inference**: The MLLM performs M separate inference steps, processing one K-slice stack at a time.

This fixed-size window approach ensures that each finding is analyzed with sufficient volumetric context while avoiding long-sequence noise and hardware memory saturation.

Step III: Final Validation and Report Generation We employ a two-phase strategy to ensure clinical accuracy and coherence. To prevent redundancy and hallucinations inherent in direct concatenation, an MLLM-based "Reflector" first audits the localized interpretations from Step II. Subsequently, the verified findings are synthesized into a standardized radiology report, structured into professional "Findings" and "Impression" sections.

3.3 MediReflect: A Self-Reflective Mechanism for MLLM Evolution

We propose ***MediReflect*** (Figure 2), a **medi**cal application–oriented **reflect**ive learning mechanism, which serves as an adaptive framework that mimics clinical reflective reasoning through two core components: multi-dimensional loss, and trajectory-based knowledge retrieval.

Component I: Multi-Dimensional Penalty Loss *MediReflect* uses a composite loss:

$$\mathcal{L}_{\text{total}} = \omega_1 \mathcal{L}_{\text{slice}} + \omega_2 \mathcal{L}_{\text{report}} + \omega_3 \mathcal{L}_{\text{disease}}, \quad (3)$$

where ω_i denotes the dynamic weight of each objective. The three components are:

1. **Slice Anomaly Loss** ($\mathcal{L}_{\text{slice}}$): Binary cross-entropy over N slices:

$$\mathcal{L}_{\text{slice}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[Y^i_{\text{slice}} \log(P^i_{\text{slice}}) + (1 - Y^i_{\text{slice}}) \log(1 - P^i_{\text{slice}}) \Big]. \quad (4)$$

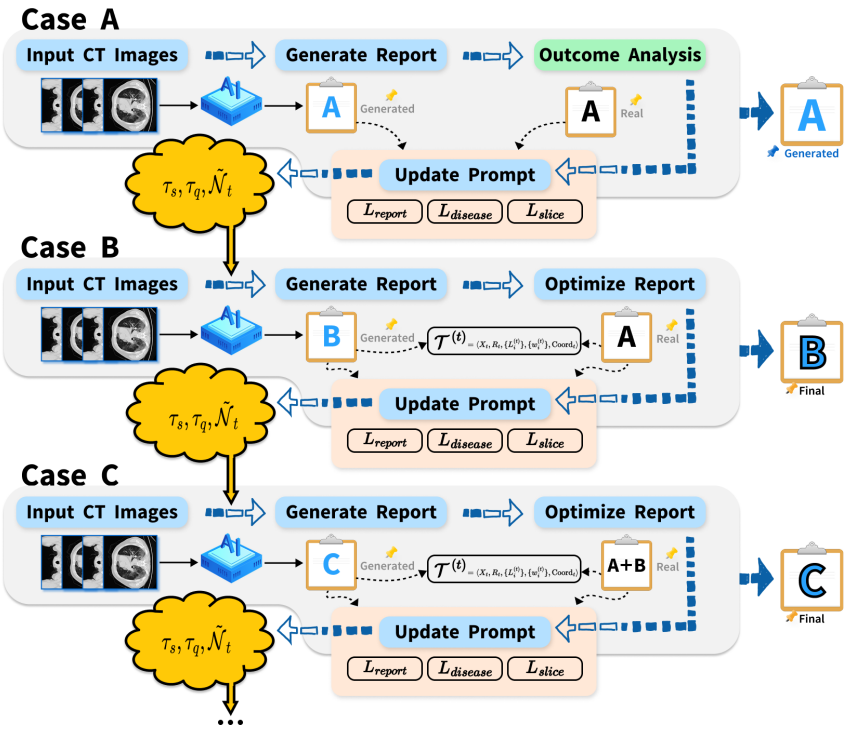


Figure 2: The MediReflect mechanism: an iterative process that alternates between reflection on visual input and retrieval of clinical knowledge to refine diagnostic reasoning.

Here, $Y^i_{\text{slice}} \in \{0, 1\}$ indicates whether the i-th slice contains a lesion, and P^i_{slice} is the predicted anomaly probability from the screening module.

2. **Disease Classification Loss** ($\mathcal{L}_{\text{disease}}$): Multi-label binary cross-entropy over M diagnostic categories:

$$\mathcal{L}_{\text{disease}} = -\frac{1}{M}\sum_{k=1}^{M}\Big[Y^k_{\text{disease}}\log(P^k_{\text{disease}}) + (1 - Y^k_{\text{disease}})\log(1 - P^k_{\text{disease}})\Big]. \tag{5}$$

Here, $Y^k_{\text{disease}} \in \{0, 1\}$ indicates the presence of the k-th disease in the ground truth, and P^k_{disease} is the predicted probability for that category.

3. **Report Quality Loss** ($\mathcal{L}_{\text{report}}$):

$$\mathcal{L}_{\text{report}} = 1 - (\lambda_{\text{sem}} \cdot \text{sim}_k + \lambda_{\text{str}} \cdot \text{comp}_s)\ . \tag{6}$$

We set $\lambda_{\text{sem}} = \lambda_{\text{str}} = 0.5$, where:

- sim_k: Jaccard similarity between clinical entities extracted by **RadGraph** (Jain et al., 2021) from the generated report ($\mathcal{E}_{\text{gen}}$) and the ground-truth report ($\mathcal{E}_{\text{gt}}$):

$$\text{sim}_k = \frac{|\mathcal{E}_{\text{gen}} \cap \mathcal{E}_{\text{gt}}|}{|\mathcal{E}_{\text{gen}} \cup \mathcal{E}_{\text{gt}}| + \epsilon},$$

where ϵ is a small constant.

- comp_s: Structural completeness ratio:

$$\text{comp}_s = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\mathbb{I}(s \in R_{\text{gen}}),$$

where $\mathcal{S} = \{\text{“Findings”}, \text{“Impression”}\}$ is the set of required sections, R_{gen} is the generated report, and $\mathbb{I}(\cdot) = 1$ if section header s appears in R_{gen}, and 0 otherwise.

Component II: Trajectory Learning and Knowledge Retrieval *MediReflect* logs the evolution of each training case into a structured trajectory $T^{(t)}$, forming a searchable knowledge base $\mathcal{K}$:

$$T^{(t)} = \Big\langle X_t, R^{\text{gen}}_t, R^{\text{gt}}_t, Y^{\text{slice}}_t, Y^{\text{disease}}_t, \{L^{(t)}_i\}, \{w^{(t)}_i\}, P^t_{\text{slice}}\Big\rangle\ . \tag{7}$$

Here, X_t denotes the input CT volume; R^{gen}_t and R^{gt}_t represent the generated and ground-truth reports; $Y^{(\cdot)}_t$ are the ground-truth labels; $\{L^{(t)}_i\}$ and

$\{w_i^{(t)}\}$ store the historical loss values and dynamic weights; and P_{slice}^t is the stored anomaly probability vector generated during the training phase.

During inference, since ground-truth profiles are unavailable, we construct a **Predicted Spatial Anomaly Profile** $\hat{P} = [\hat{y}_1, \ldots, \hat{y}_N]$, where $\hat{y}_i$ is the anomaly confidence score of the i-th slice obtained from the Stage-1 Screen-CLIP. We then perform KNN retrieval to identify a set of clinically similar historical cases $\mathcal{N}_X$:

$$\mathcal{N}_X = \text{KNN}\left(\hat{P}, \mathcal{K}, k\right). \tag{8}$$

where k is the number of neighbors retrieved from knowledge base $\mathcal{K}$. These retrieved trajectories undergo an LLM-based filtration to evaluate reference-worthiness. The selected high-quality trajectories are then utilized to assist prompt tuning and optimize the current diagnostic reasoning.

4 Results

We conduct comprehensive experiments to evaluate the proposed framework across three key steps, including rapid abnormality screening, diagnostic interpretation, and report generation quality, together with a component-level ablation study to assess the contribution of each module. For the end-to-end pipeline, we first analyze the performance of the trained CLIP-based model in the rapid abnormality screening stage and examine whether it can efficiently identify suspicious slices from large 3D volumes. We further compare our framework with state-of-the-art 3D MLLMs such as M3D-LaMed-Phi-3-4B (Shi et al., 2025) and RADFM (Wu et al., 2025) to evaluate diagnostic interpretation performance and the quality of generated clinical reports. Finally, a comprehensive ablation study is conducted to quantify the contribution of each individual component to the overall performance.

Unless otherwise specified, all comparative experiments adopt a unified prompting strategy and experimental setup, with Qwen3-VL-8B-Instruct serving as the core MLLM. Detailed implementation settings and prompt templates are provided in Tables S2–S4.

4.1 Evaluation of Initial Rapid Abnormality Screening (on Step I)

Given the large volume of slices in 3D CT data and the sparsely distributed nature of clinically relevant abnormalities, an effective and efficient screening mechanism is essential to enable scalable downstream diagnostic reasoning. We evaluated the screening performance of our framework on a patient-level 7:3 split of the doctor-annotated subset to ensure no data leakage across subjects. Our **Screen-CLIP** was benchmarked against state-of-the-art MLLMs (**Qwen2.5-VL-72B** (Bai et al., 2025) and **GPT-4o** (Hurst et al., 2024)) in a zero-shot setting, and standard supervised backbones (**PMC-CLIP** (Lin et al., 2023), **ViT-Base**, and **ResNet-50**).

Method	Accuracy	Precision	Recall	F1-Score	AUC
Zero-shot MLLMs					
Qwen2.5-VL-72B	0.62	0.57	0.69	0.63	-
GPT-4o	0.68	**0.89**	0.35	0.50	-
Supervised Models					
PMC-CLIP	0.56	0.52	0.45	0.49	0.60
ViT-Base	0.77	0.77	0.71	0.74	0.84
ResNet-50	0.71	0.69	0.67	0.68	0.79
Screen-CLIP (Ours)	**0.80**	0.80	**0.83**	**0.82**	**0.87**

Table 1: Quantitative comparison of abnormality screening performance on the test set. The best results are highlighted in **bold**. For zero-shot MLLMs, AUC is not reported due to their discrete text-based outputs.

As summarized in Table 1, **Screen-CLIP** significantly outperforms all baseline models, achieving the highest **Accuracy (0.80)**, **F1-score (0.82)**, and **AUC (0.87)**. A pivotal observation is the trade-off between Precision and Recall: while **GPT-4o** yields a high Precision (0.890), its extremely low Recall (0.35) indicates an overly conservative diagnostic bias, which is clinically unacceptable for a screening task due to the heightened risk of missed diagnoses. In contrast, **Screen-CLIP** maintains a robust **Recall (0.83)** while preserving high precision, effectively minimizing false negatives. Furthermore, our model surpasses the best-performing supervised baseline (**ViT-Base**) by **3.6% in AUC** and **16.9% in Recall**. This reliable screening performance serves as a critical gatekeeper for the subsequent spatial anomaly retrieval and diagnostic interpretation stages.

4.2 Comparative Evaluation of Medical Diagnostic Capability (on Step II)

To assess whether the proposed framework achieves competitive diagnostic reasoning performance beyond initial screening, we conduct a controlled comparison against representative state-of-the-art 3D medical MLLMs.

In the M3D (Bai et al., 2024) and RadFM (Wu et al., 2025) model testing, we load the pretrained

M3D model and RadFM on an A800 GPU separately. In the image preprocessing section, the `nibabel` library is used to load NIfTI-format medical imaging files and extract the image data. Next, the image is resampled to meet the model's input requirements, and the output report is generated in the specified format. In this process, we provide the same prompt to M3D, RadFM, and ours; the details can be seen in Table S2, Table S3, and Table S4. Results are shown in Table 2. Our model demonstrates strong disease recognition capabilities, achieving a higher F1-score on 5 out of 8 diseases compared to M3D and RadFM, including significant improvements in categories like Atelectasis and Cardiomegaly.

Disease Feature	M3D citeref18	RadFM (Wu et al., 2025)	Ours
Arterial wall calcification	0.27	0.13	**0.29**
Cardiomegaly	0.09	0.09	**0.17**
Atelectasis	0.34	0.29	**0.53**
Lung nodule	0.29	**0.42**	0.40
Lung opacity	0.17	**0.72**	0.67
Pleural effusion	0.11	0.08	**0.17**
Consolidation	0.37	0.32	**0.36**
Interlobular septal thickening	0.43	**0.67**	0.40
Macro-average	0.26	0.34	**0.37**

Table 2: Comparison of **F1 Score** for Disease Feature Classification between different models. The F1-Score value of some diseases is 0 because the sample is rare or difficult to distinguish.

4.3 Quality Assessment of Medical Report Generation (on Step III)

To evaluate whether the proposed framework can translate diagnostic understanding into coherent and clinically meaningful reports, we assess its report generation quality on a shared test set. We required the MLLM to read CT images (providing 3D CT images directly to M3D and RadFM) and generate reports on the same test set. Therefore, in this experiment, we had medical personnel select several diseases that are recognizable at the current stage. The quality of the generated reports was compared against the ground truth reports using four metrics: BLEU-Avg, BERT-Score, ROUGE-L, and METEOR. As shown in Table 3, our model achieves substantial improvements over M3D and RADFM on all metrics, achieving a significant lead over all baselines. Specifically, it improves the BLEU-Avg by 19.2% and METEOR by 4.3 points compared to the strongest baseline, M3D.

In terms of the 1–10 subjective Doctors' Score, assessed by three individuals with professional medical backgrounds, our framework obtained the highest average score of 6.16, compared with 5.89 for M3D and 5.47 for RADFM, as shown in Figure 3. This result provides preliminary evidence of improved perceived report quality.

Model	BLEU Avg	ROUGE-L	METEOR	BERT Score	Doctors' Score
Ours	29.03	16.73	23.78	82.29	6.16
M3D	24.35	13.38	19.48	81.89	5.89
RADFM	11.71	9.69	13.46	79.08	5.47

Table 3: Quantitative comparison of report generation quality across different models using automatic metrics (BLEU-Avg, ROUGE-L, METEOR, BERT-Score) and human evaluation (Doctors' Score).

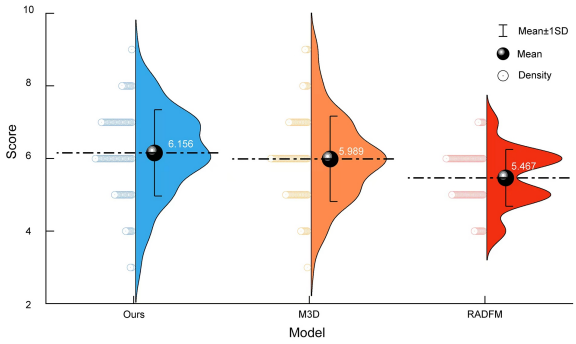


Figure 3: Distribution of doctor-assigned scores evaluating the quality and clinical applicability of model-generated medical reports. Our method achieves a higher mean score compared to other 3D medical MLLMs.

4.4 Ablation Analysis of Core Framework Components

We conduct ablation studies to evaluate the contribution of individual framework components by selectively removing Screen-CLIP, Neighbor Window, Trajectory Retrieval, and TextGrad (Reflection).

As shown in Table 4, **Screen-CLIP** is central to both efficiency and accuracy. By filtering non-informative slices, it reduces average inference latency by 61.3% (from 9.83 s to 3.80 s) and GPU memory consumption by 9.4%.

Configuration	**Latency (s)**	**GPU Memory (GB)**
Without screen-clip	9.83	19.06
With screen-clip	3.80	17.27

Table 4: Impact of screen-clip on latency and GPU memory usage.

Figure 4 further confirms its importance: remov-

ing Screen-CLIP causes a 39.7% drop in F1-score, while integrating it substantially improves throughput. Excluding the **Neighbor Window** decreases F1-score by 32.4%, indicating the importance of context for volumetric interpretation. Removing **TextGrad** reduces BLEU by 23.2%, showing its contribution to report quality. Finally, without **Trajectory Retrieval**, performance declines consistently by approximately 14%–16% across evaluation metrics.

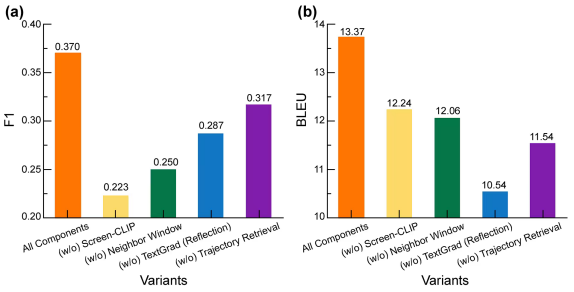


Figure 4: Component-wise ablation study. We evaluate the contribution of Screen-CLIP, Neighbor Window, Trajectory Retrieval, and TextGrad to F1 (accuracy), BLEU (quality), and latency. Removing Screen-CLIP causes the most severe degradation, with a 39.7% drop in F1.

Specifically, in the configuration without Screen-Clip, we bypassed the first-stage screening; instead, clips were input into the second-stage analysis in batches based on the maximum threshold. In the absence of Neighbor Window, rather than selecting the number of clips based on severity calculated via softmax, we solely input the clips identified as abnormal by the Screen-Clip module. For the removal of Trajectory Retrieval, the model was required to make independent judgments without referencing retrieval cases. Finally, in the experiment without TextGrad (Reflection), prompt updates were disabled.

4.5 Discussion

Although our framework achieved a solid **0.11 boost** in macro-average F1-score over M3D, the picture becomes more nuanced at the individual-disease level. This variation largely reflects the intrinsic imaging characteristics of different pathologies.

Our approach was particularly effective for volumetric lesions with clear boundaries, such as **Atelectasis (0.53)** and **Lung Nodules (0.40)**. This is consistent with our design: by mimicking how a radiologist scrolls through the Z-axis, the *MediReflect* mechanism captures spatial continuity that single-slice 2D proxies often miss. In effect, the model learned to distinguish a 3D nodule from a 2D vessel cross-section by considering the "whole picture."

However, there is a trade-off for subtle textures. We observed a slight performance drop in detecting **Consolidation** and **Interlobular Septal Thickening** compared with some baselines. These conditions are difficult because they depend on subtle textural changes that can be confused with normal pulmonary vasculature. Our *Screen-CLIP* module, while efficiently filtering healthy slices, may be somewhat too aggressive here, likely discarding slices containing faint, "threshold-level" signals before the MLLM could analyze them. This suggests that while the screening strategy works well for focal lesions, the filtering threshold may need adjustment for diffuse interstitial diseases.

Finally, like many deep learning systems, we still face the long-tail distribution problem. Although the model performed competitively on critical conditions such as **Cardiomegaly**, its generalization to extremely rare pathologies remains constrained by data scarcity. Even powerful MLLMs appear to struggle to hallucinate correct diagnoses for rare diseases without sufficient few-shot clinical examples.

In short, while our framework successfully replicates radiologist reasoning for structural abnormalities, future work should improve sensitivity to ultra-fine textural changes and robustness to class imbalance.

5 Conclusion

This study proposes a radiologist-imitation framework integrating MediReflect for closed-loop 3D CT diagnosis and reporting, with three clinical advantages. First, it improves diagnostic safety: trained on 12,000 annotated slices, Screen-CLIP achieves a recall of 0.83 and F1-score of 0.82, outperforming baselines and reducing missed diagnoses of subtle lesions. Second, it shows promising perceived report quality in a preliminary blinded human evaluation, obtaining an average Doctors' Score of 6.16 compared with 5.47–5.89 for the baselines. Third, it supports practical deployment: optimized inference reduces GPU memory cost by 9.4% per case, improving diagnostic efficiency in resource-constrained settings.

Limitations

This study has several limitations. The current experiments are mainly conducted on chest CT data, so the generalizability of the framework to other imaging modalities, anatomical regions, and multi-organ diagnostic scenarios remains to be further validated. Although Screen-CLIP improves inference efficiency, further optimization is still needed for real-time clinical deployment. In addition, the current framework does not explicitly model diagnostic uncertainty, which is important for safety-critical medical applications.

The human evaluation should be regarded as preliminary, as it involves a limited number of evaluators and does not yet include inter-rater agreement or statistical significance testing. Future work will conduct larger-scale reader studies and further improve the framework in terms of multimodal adaptation, inference efficiency, and uncertainty-aware diagnosis.

References

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, and 1 others. 2024a. Masam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024b. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. 2024. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical image analysis*, 92:103046.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, and 1 others. 2026. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, and 1 others. 2022. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. 2025. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *arXiv preprint arXiv:2503.18135*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

NT Huynh, E Zhang, O Francies, F Kuklis, T Allen, J Zhu, O Abeyakoon, F Lucka, M Betcke, J Jaros, and 1 others. 2025. A fast all-optical 3d photoacoustic scanner for clinical vascular imaging. *Nature Biomedical Engineering*, 9(5):638–655.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Charles Leduc, Frank Detterbeck, James Huang, Anja C Roden, and Moishe Liberman. 2025. Managing thymic malignancies in 2025: Migrating from masaoka to tnm. *Journal of Thoracic Oncology*, 20(12):1760–1762.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Domenico Mastrodicasa, Marly Van Assen, Merel Huisman, Tim Leiner, Eric E Williamson, Edward D Nicol, Bradley D Allen, Luca Saba, Rozemarijn Vliegenthart, and Kate Hanneman. 2025. Use of ai in cardiac ct and mri: a scientific statement from the escr, eusomii, nasci, scct, scmr, siim, and rsna. *Radiology*, 314(1):e240516.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.

Chuang Niu, Qing Lyu, Christopher D Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K Kalra, Christopher T Whitlow, and Ge Wang. 2025. Medical multimodal multitask foundation model for lung cancer screening. *Nature Communications*, 16(1):1523.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. 2022. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Philippe Raffy, Jean-François Pambrun, Ashish Kumar, David Dubois, Jay Waldron Patti, Robyn Alexandra Cairns, and Ryan Young. 2023. Deep learning body region classification of mri and ct examinations. *Journal of Digital Imaging*, 36(4):1291–1301.

Gat Rauner, Piyush B Gupta, and Charlotte Kuperwasser. 2025. From 2d to 3d and beyond: the evolution and impact of in vitro tumor models in cancer research. *Nature Methods*, 22(9):1776–1787.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1):222.

Heinz-Peter Schlemmer. 2023. The cancer epidemic : Global significance of cancer and the situation in oncological imaging. *Radiologie (Heidelb.)*, 63(1):49–56.

Yiming Shi, Xun Zhu, Kaiwen Wang, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. 2025. Med-2e3: A 2d-enhanced 3d medical multimodal large language model. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2754–2759. IEEE.

Guoan Wang, Jin Ye, Junlong Cheng, Tianbin Li, Zhaolin Chen, Jianfei Cai, Junjun He, and Bohan Zhuang. 2024a. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer.

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, and 1 others. 2025. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31(2):609–617.

Yuli Wang, Jian Peng, Yuwei Dai, Craig Jones, Haris Sair, Jinglai Shen, Nicolas Loizou, Jing Wu, Wen-Chi Hsu, Maliha Imami, and 1 others. 2024b. Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. *Advances in Neural Information Processing Systems*, 37:99947–99964.

Zhaoqing Wang, Tianqing Wan, Sijie Ma, and Yang Chai. 2024c. Multidimensional vision sensors for information processing. *Nature Nanotechnology*, 19(7):919–930.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Kaidi Zhang, Zimo Zhao, Yan Hu, and Thuy Le. 2026. Ai-driven feature selection using only survey variable descriptions: Large language models identify adolescent vaping predictors. *medRxiv*, pages 2026–03.

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

Debmalya Pal
University of California, San Diego
Department of Computer Science and Engineering
d2pal@ucsd.edu

Abstract

Biomedical language models can generate overly confident clinical statements despite incomplete or ambiguous evidence. We study whether linguistic uncertainty (the hedged epistemic stance expressed in phrases such as "consistent with" or "cannot exclude") is encoded in model representations and can be controlled without retraining. Across six biomedical language models spanning two architectures (causal decoders and bidirectional encoders), we show that uncertainty is captured by robust low-dimensional linear structure in hidden states. We then apply activation steering to manipulate this representation directly, increasing hedged generation in decoder models and inducing targeted uncertainty-related shifts in encoder representations. Together, these results show that epistemic stance is not merely a surface linguistic phenomenon but an interpretable and controllable feature of biomedical language model representations, with implications for safer and more calibrated clinical text generation.

1 Introduction

Biomedical language models are increasingly proposed as components of clinical decision support systems, from radiology report generation to literature-assisted diagnosis. Yet a gap persists between benchmark performance and clinical deployability: language models trained on biomedical text tend to produce assertions in a confident declarative register, even when the clinical evidence is incomplete. Prior work has documented this tendency in the factual domain (Singhal et al., 2023; Umapathi et al., 2023) and in models' failure to express calibrated uncertainty in their outputs (Xiong et al., 2024); the corresponding problem for *linguistic hedging style* in clinical text is the focus of this work.

Throughout this work, we use *uncertainty* to refer to the *linguistic expression* of epistemic hedging in clinical text (the use of modal constructions and speculative cues that mark epistemic stance toward a proposition), rather than to a model's confidence calibration over factual claims (Geng et al., 2024). Expressing this uncertainty ("findings are consistent with," "cannot exclude") is a learned professional skill with real safety consequences (Panicek and Hricak, 2016): confident model outputs invite action without confirmation, while hedged outputs prompt consideration of alternatives.

Existing behavioral control methods operate either at the input level (prompting, brittle (Zhao et al., 2021; Lu et al., 2022)) or the parameter level (fine-tuning, expensive). A third approach from mechanistic interpretability, *activation steering* (Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2024), intervenes directly on hidden states at inference time along directions encoding target concepts, requiring no retraining. Whether this prospect is realistic for clinical uncertainty in biomedical models is an open empirical question.

Two conditions must hold: uncertainty must be linearly encoded in hidden states, and steering along this direction must shift outputs without collapsing other generation properties. We investigate both systematically across six biomedical language models spanning two architectural families (causal decoders and bidirectional encoders), parameter scales ranging from 110M to 2.7B, and five pretraining corpora. This multi-model scope is deliberate: a finding that holds only for a single checkpoint could reflect checkpoint-specific memorization of speculation cues; establishing it across architectures, scales, and corpora is evidence that linear uncertainty encoding is a property of biomedical LM training rather than an artifact. Our five contributions are:

Code: github.com/DebmalyaPal/Probing-and-Steering-Uncertainty-in-Biomedical-Language-Models

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1066–1079
July 3-4, 2026 ©2026 Association for Computational Linguistics

- We show uncertainty is linearly decodable across all six models (80–87% 5-fold CV accuracy), establishing it as a general property of biomedical LMs rather than a model-specific artifact.
- We document a 15-fold Cohen's d improvement for mean vs. last-token pooling in BioGPT (1.33 vs. 0.09), with implications for probing biomedical LMs broadly.
- We introduce a principled two-stage layer selection (2/3-depth anchor ± 2 window) validated empirically: gains range from $+0.25$pp to $+2.50$pp, with BioBERT showing the most striking improvement (L10 over L8 anchor).
- We show that dynamic hidden-norm calibration prevents generation quality collapse across the full tested steering range for decoder models, and characterize qualitatively distinct steering regimes for BioGPT and BioMedLM arising from differences in scale and hidden dimension.
- We introduce a five-metric encoder representation-steering suite (projection shift, probability shift, flip rate, centroid cosine shift, on-axis fraction) and demonstrate consistent, architecturally uniform shifts across all four encoder models despite distinct pretraining corpora.

2 Related Work

Uncertainty in biomedical text. The BioScope corpus (Vincze et al., 2008) provides sentence-level speculation annotations used as ground truth in our work. Prior classification work (Agarwal and Yu, 2010; Adel and Schütze, 2015) treated hedging as a surface-level feature; we ask whether it constitutes an *internal representational property* of biomedical LMs. Recent clinical LLM calibration work (Umapathi et al., 2023; Singhal et al., 2023) focuses on factual accuracy rather than the expression of epistemic stance, leaving our question largely unaddressed.

Probing representations. Probing (training classifiers on frozen hidden states) has established a layered hierarchy of linguistic information in LMs (Hewitt and Manning, 2019; Tenney et al., 2019; Belinkov, 2022). In biomedical models, probing has focused on domain knowledge (Sung et al., 2021; Vulić et al., 2020); pragmatic and stance-like properties are understudied. A key caveat is that decodability does not imply task relevance (Ravichander et al., 2021), a distinction directly relevant to our steering results.

Activation steering. The linear representation hypothesis (Park et al., 2024) underlies activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), and contrastive activation addition (Panickssery et al., 2024). Methodological concerns include surface-feature confounds (Elazar et al., 2021) and out-of-distribution degradation at large perturbation magnitudes (Zou et al., 2023). Application to biomedical models is, to our knowledge, unexplored. We address this gap across six models with two architectural families and full multi-metric evaluation.

Multi-model comparative probing. Most probing studies target a single model or checkpoint (Tenney et al., 2019; Hewitt and Manning, 2019); systematic comparison across architectures remains less common. Notable exceptions include Vulić et al. (2020), who compare BERT variants on lexical knowledge, and Belinkov (2022), who survey cross-model probing methodology. Our study extends this comparative lens to a pragmatic property (epistemic stance) across two architectural families with different training objectives, examining both where uncertainty representations form and whether they support controlled steering.

3 Methods

The methodology has two stages: (i) train linear probes on frozen hidden states to identify the layer and direction $\mathbf{w}_{\hat{\ell}}$ that best separates uncertain from certain sentences (§§3.3–3.4); (ii) inject a scaled multiple of that direction at inference time via a forward hook, with no retraining (§§3.5–3.6).

3.1 Dataset

We use the BioScope corpus (Vincze et al., 2008), which provides token-level scope annotations for speculation and negation. We collapse these to sentence-level binary labels: sentences containing any speculation cue are treated as *uncertain*, all others as *certain*. This simplification discards scope information but yields a clean binary contrast set aligned with our probing objective. Parsing all three sub-corpora (abstracts, full papers, clinical reports) yields 3,065 uncertain and 13,007 certain sentences. After restricting to sentences of 30–200 characters, we sample 200 of each class (seed 42) to form a balanced 400-sentence contrast set $\mathcal{D}$. Files are enumerated alphabetically before sampling for cross-platform reproducibility. Despite length filtering, a small residual difference remains

Table 1: Models evaluated (float32, no quantization).

Model	Arch.	L	Params	Pretrain
BioGPT	Dec.	24	345M	PubMed
BioMedLM	Dec.	32	2.7B	PubMed
BioBERT	Enc.	12	110M	PubMed+PMC
ClinicalBERT	Enc.	12	110M	MIMIC-III
BlueBERT	Enc.	12	110M	PubMed+MIMIC
SciBERT	Enc.	12	110M	Broad sci.

(27.1 vs. 24.6 tokens on average), which we later control via orthogonalization.

3.2 Models and representation extraction

We study six biomedical LMs across two families (Table 1): two causal decoders, BioGPT (Luo et al., 2022) and BioMedLM (Bolton et al., 2024), and four bidirectional encoders, BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), and SciBERT (Beltagy et al., 2019). For sentence s tokenized into T tokens, we extract mean-pooled hidden states at layer ℓ:

$$\mathbf{h}_\ell(s) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{H}_\ell(s)[t,:] \tag{1}$$

We use mean pooling uniformly; Section 4.1 shows it yields Cohen's $d = 1.33$ vs. 0.09 for last-token pooling, motivating this choice.

3.3 Linear probing

For each layer ℓ, we train a logistic regression probe:

$$p(y{=}\text{unc.} \mid \mathbf{h}_\ell(s)) = \sigma\Big(\mathbf{w}_\ell^\top \mathbf{h}_\ell(s) + b_\ell\Big) \tag{2}$$

with ℓ_2-regularization $C = 0.1$. Evaluation uses stratified 5-fold CV on the 400-sentence set; we report mean accuracy, std, and 95% CI.

3.4 Principled layer selection

Stage 1: 2/3-depth anchor.

$$\ell^* = \text{round}\big(\tfrac{2N}{3}\big) \tag{3}$$

This heuristic is grounded in probing studies showing that middle-to-upper layers encode the most abstract linguistic properties (Tenney et al., 2019; Belinkov, 2022). For decoders it also balances representational maturity against propagation distance (enough remaining blocks for the injected signal to influence generation). The formula gives $\ell^* = 8$ for all 12-layer BERT variants and $\ell^* = 16$ for BioGPT; the latter coincides with the layer examined in Luo et al. (2022).

Table 2: Steering layer selection. $\ell^* = \text{round}(2N/3)$; $\hat{\ell}$ = window-best (Eq. 4). † = window differs from anchor.

Model	N	ℓ^*	Acc@ℓ^*	$\hat{\ell}$	Acc@$\hat{\ell}$	Std
BioGPT	24	16	85.50%	16	85.50%	2.44%
BioMedLM†	32	21	84.00%	22	84.25%	1.90%
BioBERT†	12	8	84.00%	10	86.50%	2.40%
ClinicalBERT†	12	8	82.00%	9	82.75%	3.79%
BlueBERT	12	8	80.75%	8	80.75%	5.35%
SciBERT†	12	8	84.75%	10	85.00%	3.42%

Stage 2: Local empirical refinement. Probe accuracy curves are locally flat across several adjacent layers. We search within ± 2 layers of the anchor, selecting by highest mean accuracy (ties broken by minimum std):

$$\hat{\ell} = \underset{\ell \in [\ell^*-2,\, \ell^*+2] \cap \{1,\ldots,N\}}{\text{argmax}} \big(\bar{a}_\ell,\ -s_\ell\big) \tag{4}$$

where $\bar{a}_\ell$ is the mean CV accuracy and s_ℓ the standard deviation at layer ℓ. The tuple $(\bar{a}_\ell, -s_\ell)$ encodes lexicographic tie-breaking: select the layer with highest mean accuracy; break ties by lowest standard deviation. Layer 0 (embedding output) is excluded as it is not a hookable transformer-block module. Table 2 and Figure 2 show that the window correctly identifies BioBERT L10 over the L8 anchor (+2.50pp); BioGPT and BlueBERT show zero gain (anchor = window-selected layer), and the remaining three models (BioMedLM, ClinicalBERT, SciBERT) gain between +0.25pp and +0.75pp.

3.5 Hidden-norm calibration and steering directions

Steering magnitude is parameterized fractionally:

$$\alpha_{\text{eff}} = \alpha_{\text{frac}} \cdot \|\mathbf{h}_{\hat{\ell}}\|_{\text{est}} \tag{5}$$

where $\|\mathbf{h}_{\hat{\ell}}\|_{\text{est}}$ is estimated dynamically from 8 generic biomedical sentences. This makes α_{frac} interpretable as a fraction of typical hidden-state magnitude across layers and models. For BioGPT L16, dynamic calibration yields ≈ 205, versus the static value 616 used in our uncalibrated baseline (derived from a single prompt without averaging), a $3\times$ discrepancy that produced perplexity collapse (> 400) at $\alpha_{\text{frac}} = 0.25$; calibrated norms prevent this entirely.

We construct two steering directions at $\hat{\ell}$, both derived from the probe in Eq. 2. For the steering direction the probe is retrained on all 400 sentences (rather than per-fold subsets) to obtain a maximally

stable direction; Finding 3 shows this recovers per-fold CV accuracy completely.

The **probe direction** is:

$$\mathbf{v}_{\text{probe}} = \mathbf{w}_{\hat{\ell}}^{\text{full}} \tag{6}$$

The **orthogonalized direction** removes two candidate confound directions from $\mathbf{v}_{\text{probe}}$: a *length direction* $\mathbf{v}_{\text{len}}$ (regression weights predicting token count) and a *hedge-count direction* $\mathbf{v}_{\text{hedge}}$ (regression weights predicting hedging-vocabulary occurrences). We define the orthogonalization operator as:

$$\text{orth}(\mathbf{u}, \mathbf{v}) = \mathbf{u} - \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v} \tag{7}$$

and apply it sequentially:

$$\mathbf{v}_{\text{ortho}} = \text{orth}\big(\text{orth}(\mathbf{v}_{\text{probe}},\ \mathbf{v}_{\text{len}}),\ \mathbf{v}_{\text{hedge}}\big) \tag{8}$$

The resulting direction is orthogonal to both confounds while retaining maximum alignment with the original probe direction. For BioGPT L16, $\cos(\mathbf{v}_{\text{probe}}, \mathbf{v}_{\text{ortho}}) = 0.95$, confirming that orthogonalization makes only a small angular adjustment. Note that this procedure controls for hedge-count as a scalar surface feature; it does not fully separate lexical from semantic content, since the probe direction itself is derived from annotations defined partly by the presence of hedging vocabulary.

3.6 Steering protocols

Decoder steering. We register a forward hook at $\hat{\ell}$ that adds $\alpha_{\text{eff}} \cdot \hat{\mathbf{v}}$ to all token positions during both prefill and each autoregressive decoding step. We sweep $\alpha_{\text{frac}} \in \{0, 0.025, \ldots, 0.25\}$ (9 values) for probe and ortho directions; 20 seeds $\times$ 5 radiology prompts $=$ 100 samples per configuration. BioMedLM uses a coarser step of 0.05 (6 values) due to its larger memory footprint; the monotonically increasing response gives no indication that finer resolution would alter the qualitative findings. Outputs are scored on hedge score, lexical diversity, token validity, and perplexity (see Appendix E).

Encoder steering. For encoder models (no generation), we inject $\alpha_{\text{eff}} \cdot \hat{\mathbf{v}}$ at $\hat{\ell}$ and compare modified vs. original mean-pooled hidden states across 80 held-out test sentences. We compute five metrics: projection shift Δ_{proj}, probability shift Δ_{prob}, decision-boundary flip rate, centroid cosine shift $\Delta_{\cos}$, and on-axis fraction (formal definitions in Appendix E).

Table 3: Probe CV accuracy at selected steering layer $\hat{\ell}$.

Model	$\hat{\ell}$	L0 acc	Acc@$\hat{\ell}$	Std
BioGPT	16	82.00%	85.50%	2.44%
BioMedLM	22	55.25%	84.25%	1.90%
BioBERT	10	76.00%	86.50%	2.40%
ClinicalBERT	9	77.75%	82.75%	3.79%
BlueBERT	8	77.00%	80.75%	5.35%
SciBERT	10	80.50%	85.00%	3.42%

Compute. All canonical experiments use Python 3.11.9, PyTorch 2.5.1+cu121, Transformers 4.49.0, NVIDIA GPU, CUDA 12.1, float32. bfloat16/float16 produce equivalent probe accuracies on the same CUDA device (within 0.25pp). Platform-level non-determinism (CUDA vs. MPS) can shift layer selection within the ± 2 window; CUDA results are canonical.

4 Results

4.1 Pooling method determines observable structure

At BioGPT layer 16, **last-token pooling** yields Cohen's d=0.088 (57% accuracy), while **mean pooling** yields d=1.331 (85.50% probe CV accuracy at this layer). Last-token representations are dominated by the final content token (e.g. *pneumonia*), which is informative about topic but not epistemic stance. Mean pooling surfaces distributed sentence-level properties including hedging. All subsequent analyses use mean-pooled hidden states.

4.2 Uncertainty is linearly decodable across all six models

Table 3 and Figure 1 report probe accuracy across models.

Finding 1: Robust cross-model decodability. All six models achieve 80–87% CV accuracy at the selected steering layer. The narrow spread across models with different architectures, sizes, and pretraining corpora indicates that linear encoding of epistemic stance is a general property of biomedical LMs rather than a checkpoint-specific artifact.

Finding 2: Layer-0 accuracy reveals family-level differences. BioGPT achieves 82% at layer 0; uncertainty is already largely encoded in the token embeddings through surface lexical statistics. BioMedLM achieves only 55% at layer 0 (near chance), meaning its uncertainty encoding is built entirely by transformer computation despite shar-

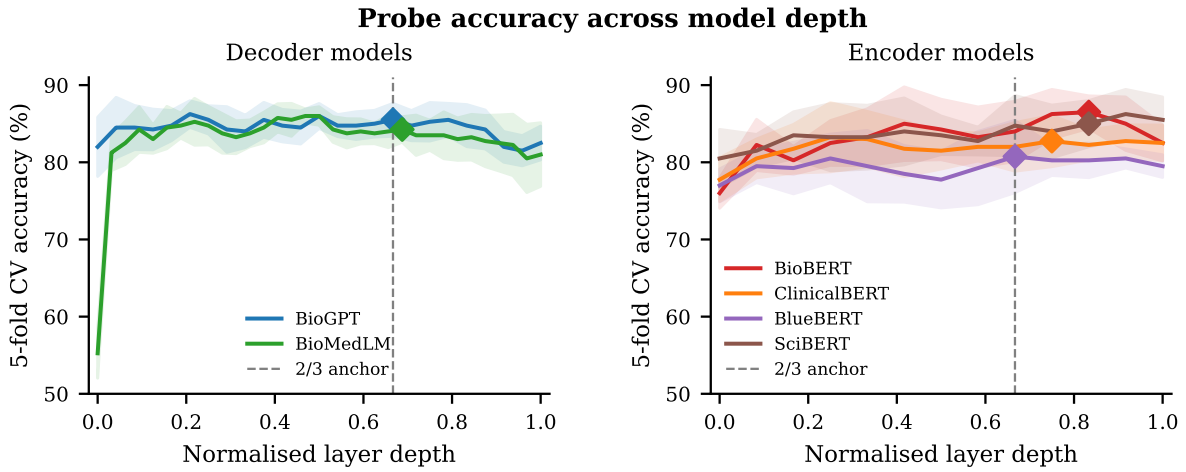


Figure 1: Probe accuracy across normalised layer depth for all models. Shaded bands are 95% CI. Diamonds mark the window-selected layer $\hat{\ell}$. Dashed line marks the 2/3-depth anchor.

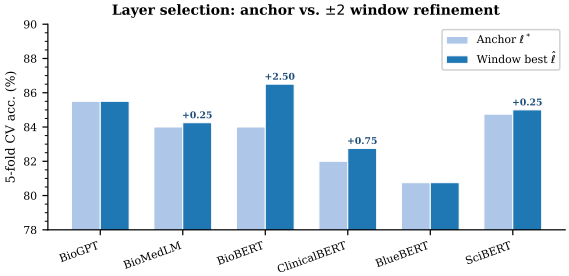


Figure 2: Anchor vs. window-best accuracy for all models. Annotations show percentage-point gain from the ±2 refinement; BioGPT and BlueBERT show zero gain (anchor = window-selected layer).

ing the same pretraining corpus as BioGPT. Encoder models occupy an intermediate position (76–81%), consistent with bidirectional attention enabling richer early contextualization even within the embedding layer's immediate context. This family-level contrast is clearly visible in Figure 1.

Finding 3: The uncertainty direction is stable and concentrated. The non-trivial question is whether the learned direction is *consistent*: does a single direction computed once on all 400 sentences, frozen across folds, recover the same accuracy as a per-fold adaptive probe free to re-orient across the full representation space? We test this by projecting onto $\mathbf{v}_{\text{probe}}$ and training only a scalar threshold per fold. For BioGPT L16, the frozen direction achieves 85.50%, matching the per-fold adaptive probe (2.44% std); encoder models recover $> 98\%$ of per-fold accuracy. This confirms that uncertainty is concentrated along a single *consistent* geometric axis. Orthogonalizing against sentence length and hedge-count reduces accuracy by < 1pp ($85.50\% \rightarrow 85.00\%$), confirming the direction encodes epistemic stance, not surface statistics. Full layer-by-layer tables appear in Appendix F.

4.3 Decoder steering: controlled range, non-monotonic peak

Figure 3 shows BioGPT decoder steering at L16 with calibrated norms ($\approx$ 205). Four findings emerge.

Finding 1: Clean steering at low-to-mid magnitudes. Hedge score rises from 0.35 (baseline) to 0.90 at $\alpha_{\text{frac}} = 0.125$ (probe direction, $2.6\times$ increase) while lexical diversity (0.94) and perplexity (20.3) remain near baseline (0.95 and 17.0). This is the signature of a clean semantic intervention.

Finding 2: Calibration prevents collapse. With static norm 616, perplexity at $\alpha_{\text{frac}} = 0.25$ exceeds 400. With calibrated norm $\approx$ 205 ($3\times$ smaller effective perturbation), perplexity peaks at 43.7 ($2.6\times$ baseline) and lexical diversity stays at or above 0.91 throughout the tested range.

Finding 3: Non-monotonic hedge response. The probe-direction hedge score peaks at $\alpha_{\text{frac}} = 0.125$ ($H = 0.90$), then oscillates between 0.80 and 0.88 above the peak ($0.84 \rightarrow 0.80 \rightarrow 0.88$ at $\alpha_{\text{frac}} \in \{0.15, 0.20, 0.25\}$), indicating the model transitions through different response regimes even within the collapse-free range. The ortho direction mirrors this with a slightly higher peak ($H = 0.97$ at $\alpha_{\text{frac}} = 0.125$), consistent with the probe direc-

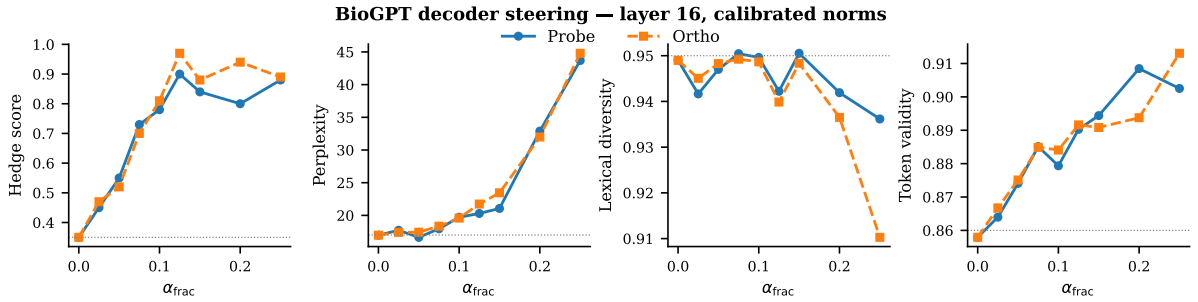


Figure 3: BioGPT decoder steering at layer 16 (calibrated norms, $\|\mathbf{h}_{16}\|_{\text{est}} \approx 205$). Dotted horizontal lines mark baseline values. Both directions show a non-monotonic hedge peak at $\alpha_{\text{frac}} = 0.125$ (Table 8) and no collapse through $\alpha_{\text{frac}} = 0.25$. The visual peak may appear at ≈ 0.10 due to marker interpolation; the numerical peak is at 0.125.

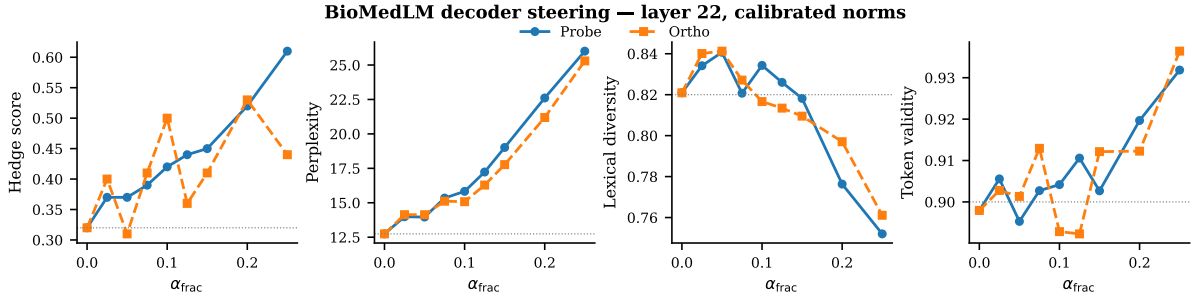


Figure 4: BioMedLM decoder steering at layer 22 (calibrated norms, $\|\mathbf{h}_{22}\|_{\text{est}} \approx 228$, $d = 2560$). Dotted horizontal lines mark baseline values. The hedge response is monotonically increasing with no collapse through $\alpha_{\text{frac}} = 0.25$, in contrast to BioGPT's non-monotonic profile (Figure 3).

tion carrying a small residual surface-feature component that the ortho direction removes. Note also that PPL dips slightly below baseline at $\alpha_{\text{frac}} = 0.05$ (16.6 vs. 17.0): at low magnitudes, the steering vector may nudge outputs closer to the PubMed training manifold (hedge vocabulary is dense there), producing a transient perplexity reduction; this effect disappears at higher magnitudes. Full tables are in Appendix G.

BioMedLM decoder steering at layer 22, calibrated norms. Figure 4 shows the BioMedLM steering profile. At BioMedLM layer 22 ($\|\mathbf{h}\|_{\text{est}} \approx 228$, $d = 2560$), three features distinguish the steering profile from BioGPT's.

Gradual monotonic response. Under the probe direction, hedge score rises from 0.32 (baseline) to 0.61 at $\alpha_{\text{frac}} = 0.25$, a near-linear $1.9\times$ increase with no mid-range dip. Perplexity climbs from 12.7 to 26.0 ($2.0\times$), and lexical diversity declines gradually ($0.82 \rightarrow 0.75$). In contrast to BioGPT, whose hedge score peaks non-monotonically at $\alpha_{\text{frac}} = 0.125$ and then dips, BioMedLM's response is monotonically increasing throughout with no generation quality collapse within the tested range.

Probe direction outperforms ortho. The orthogonalized direction achieves hedge 0.44 at $\alpha_{\text{frac}} = 0.25$ versus 0.61 for the probe direction, a reversal of the BioGPT pattern, where ortho peaked at $H = 0.97$ versus probe's $H = 0.90$. This suggests that for BioMedLM the surface-feature component of the probe direction contributes constructively to the steering effect rather than acting as a confound.

Greater distributional resilience. We attribute BioMedLM's wider collapse-free range to two factors. Its hidden dimension ($d = 2560$) is $2.5\times$ larger than BioGPT's ($d = 1024$): a fixed fractional perturbation covers a proportionally smaller fraction of the representation space per coordinate. Additionally, BioMedLM's near-chance layer-0 accuracy (55.25%) indicates that its uncertainty representation is built through transformer computation rather than surface lexical cues, yielding a steering direction that is more diffuse in activation space and less prone to out-of-distribution excursions at moderate magnitudes. Full numerical results appear in Appendix G.

4.4 Encoder representation steering

Figure 5 shows all five metrics across the four encoder models (probe direction). Key observations:

(i) Consistent positive shifts. All models show strongly positive Δ_{proj} and Δ_{prob}, confirming the steering vector reliably shifts representations toward the uncertain pole.

(ii) Flip-rate saturation. Flip rate rises rapidly to $\approx 50\%$ by $\alpha_{\text{frac}} = 0.10$ and then saturates, indicating a fixed ceiling of flippable sentences is reached once perturbation exceeds a threshold.

(iii) Centroid cosine sign reversal. $\Delta_{\cos}$ is slightly positive at small α_{frac} but turns negative above ≈ 0.10. The probe direction and the mean uncertain-centroid direction are correlated but not identical; large perturbations along the former overshoot the latter, moving representations *away* from the uncertain centroid even as probe-axis projections keep increasing.

(iv) On-axis fraction. For the probe direction, the on-axis fraction is 1.000 by construction. For the ortho direction, it falls to 0.97–0.98 (model-dependent), confirming that orthogonalization introduces only a small geometric deviation (see Appendix G).

(v) Model ordering. SciBERT shows the largest shifts ($\Delta_{\text{proj}} = 1.41$, $\Delta_{\text{prob}} = 0.40$ at $\alpha_{\text{frac}} = 0.05$); BlueBERT shows the smallest (0.77, 0.30), consistent with their relative probe accuracies.

4.5 Cross-model patterns and architectural effects

The probing and steering results together reveal several regularities that illuminate how architecture, scale, and pretraining interact with uncertainty encoding.

Layer-0 accuracy reflects pretraining objective and scale. The gap between the two decoder models at layer 0 is the sharpest contrast in our study: BioGPT achieves 82.00% while BioMedLM achieves only 55.25%, despite both being causal LMs trained on PubMed. BioGPT (345M parameters) apparently relies on surface lexical statistics (hedging tokens are correlated with their local embedding context), whereas BioMedLM (2.7B parameters) abstracts uncertainty into higher-order contextual representations that are not yet formed at layer 0. Encoder models occupy an intermediate position (76–81%), consistent with bidirectional attention enabling richer early contextualization across the full sentence. The pattern is consistent with scale (or the associated increase in hidden dimension and training compute) driving the shift from surface-dependent to computation-dependent uncertainty encoding, though the two decoders differ simultaneously in multiple factors and a controlled ablation would be needed to isolate scale as the causal variable.

Peak accuracy converges despite divergent starting points. Despite the large layer-0 variation, all six models converge to 80–87% at their selected steering layers. This convergence implies a ceiling (set either by the BioScope contrast set composition or by the intrinsic linearly accessible information in 400 sentences) that all models reach regardless of how early their uncertainty representation forms. The encoder ordering at peak accuracy (SciBERT 85.0%, BioBERT 86.5%, ClinicalBERT 82.75%, BlueBERT 80.75%) correlates with pretraining corpus breadth and biomedical specificity, suggesting that broader domain exposure slightly sharpens the uncertainty axis even when absolute accuracy differences are modest.

Encoder steering response is architecturally uniform. All four BERT-family encoders show nearly identical α_{frac} response curves for Δ_{proj} and Δ_{prob} despite being pretrained on corpora ranging from clinical notes (MIMIC-III) to broad scientific text. The decoder contrast is visible in Figure 9. The flip-rate saturation ceiling (50–54%) is consistent across all four, suggesting it reflects BioScope test-set composition rather than model-specific geometry. The model ordering on shift magnitude (SciBERT $>$ ClinBERT $>$ BioBERT $>$ BlueBERT) correlates with peak probe accuracy, confirming that a sharper uncertainty axis produces stronger representational steering effects. This uniformity has a practical implication: uncertainty steering may be transferable across BERT-family encoder variants without per-model calibration of the steering direction, provided probe accuracy is used to select the appropriate layer.

Layer selection interacts with architecture. The ± 2 window refinement matters most for encoders (BioBERT $+2.50$pp, SciBERT $+0.25$pp, ClinicalBERT $+0.75$pp) and is trivially zero for BioGPT. In BERT-family models, probe accuracy can vary non-monotonically near the 2/3 anchor as the `[CLS]` representation transitions between syntactic and semantic processing stages. For causal decoders, the layer-by-layer accuracy profile is flat

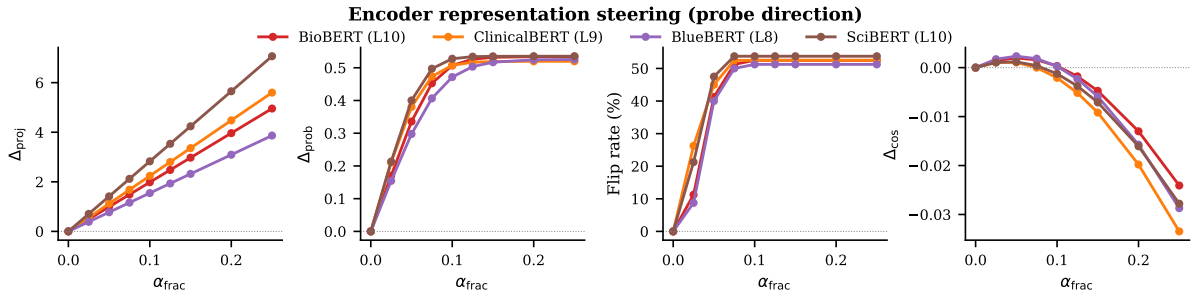


Figure 5: Encoder representation steering (probe direction) across all four models. Top-left: Δ_{proj} (all models positive, linear in α). Top-right: Δ_{prob} (saturates above $\alpha \approx 0.10$). Bottom-left: flip rate (saturates at $\approx 50\%$). Bottom-right: Δ_{cos} (sign reversal above $\alpha \approx 0.10$, dotted line marks zero).

across adjacent layers (Table 5): any layer in the ± 2 window performs equivalently, making precise selection less consequential. This asymmetry supports applying the full window search to all architectures while noting that it adds most value for bidirectional encoders.

5 Discussion

Uncertainty is cleanly decodable across all six models. The narrow range of peak accuracies (80–87%) across models with different architectures, sizes, and pretraining corpora suggests that linear encoding of epistemic stance is a general property of biomedical LMs trained on scientific or clinical text. The layer-0 contrast (BioGPT at 82% vs. BioMedLM at 55%) is particularly informative: for BioGPT, much of the signal is already present in the token embeddings through surface lexical cues, whereas BioMedLM appears to build much of the uncertainty-associated signal through transformer computation despite sharing the same pretraining corpus.

The uncertainty direction is stable but not purely semantic. Orthogonalizing against sentence length and hedge-count reduces probe accuracy by less than 1pp, which shows the representation is not reducible to these two surface statistics. However, this should not be read as evidence that the direction is purely semantic: the probe is trained on BioScope annotations that are themselves defined partly by the presence of hedging vocabulary. Orthogonalizing against a scalar hedge-count feature cannot fully decouple lexical from conceptual content. What we can say is that the direction captures more than a simple hedge-word lookup, and that this residual structure supports activation steering, but the precise mix of lexical and higher-level epistemic content remains an open question.

Layer selection is not arbitrary. The 2/3-depth anchor is a principled starting point, and the ± 2 window consistently identifies equal-or-better layers without over-searching. The BioBERT result ($+2.50$pp, L10 over L8) is the clearest demonstration: the accuracy plateau is not uniformly flat, and the window correctly finds the local optimum. The restriction to ± 2 is deliberate: wider windows risk selecting layers at qualitatively different processing stages.

Dynamic calibration is essential; decodability $\neq$ steerability. For decoders, the critical lesson is that hidden-norm calibration determines whether steering is controlled or catastrophic: a $3\times$ static over-estimate produced collapse at $\alpha_{\text{frac}} = 0.25$; dynamic calibration eliminates it across the entire tested range. Yet even controlled steering is non-monotonic: the peak hedge effect occurs at intermediate magnitudes and weakens above it, suggesting the model transitions between response regimes. For encoders, the Δ_{cos} sign reversal demonstrates that even geometrically precise steering along the probe direction does not guarantee movement toward the uncertain cluster: linear decodability and directional steerability are not equivalent properties.

PPL measurement under the unsteered model. Perplexity is scored under the *unsteered* model p_θ (Eq. 10), not the steered one, because the steered model's own likelihood is inflated along the steering direction. This provides a model-independent fluency proxy, but some PPL increase may reflect lexical style shift (more hedging vocabulary) rather than genuine quality degradation; token validity and lexical diversity serve as complementary

checks.

Multi-metric evaluation is necessary. In the uncalibrated regime (PPL > 400), hedge score alone would suggest strong success. The same logic applies to encoder steering: a large Δ_{proj} paired with negative $\Delta_{\cos}$ is ambiguous without the full metric suite. For steering interventions in safety-sensitive clinical domains, we argue that the target-behavior metric must be paired with at least one fluency/diversity metric (decoders) or one probability/boundary-crossing metric (encoders).

Decoder scale shapes the steering regime. The BioGPT–BioMedLM comparison constitutes a natural experiment: two causal LMs trained on the same corpus but differing in scale by $8\times$ (345M vs. 2.7B parameters) and hidden dimension by $2.5\times$ (1024 vs. 2560). Their steering profiles differ qualitatively. BioGPT exhibits a non-monotonic hedge peak at $\alpha_{\text{frac}} = 0.125$ with a sharp perplexity inflection above it, suggesting its representation space tightly couples the uncertainty direction to generation quality. BioMedLM exhibits a monotonic, collapse-free profile across the same range, and its ortho direction underperforms its probe direction, a reversal of the BioGPT pattern. These observations suggest that model scale modulates not merely the magnitude of the steering response but its qualitative character: larger models appear to afford wider collapse-free operating ranges, with the trade-off of lower peak hedge gains within that range. Whether this is a general scaling property or specific to the PubMed pretraining distribution remains open.

Encoder pretraining distribution has limited influence on steering geometry. The near-identical steering response curves across four BERT-family encoders pretrained on corpora ranging from clinical notes (MIMIC-III) to broad scientific text constitute our strongest evidence that linear uncertainty encoding is corpus-independent when the training objective is masked language modeling. If hedging representation were primarily a function of clinical text exposure, ClinicalBERT would be expected to show qualitatively different behavior from SciBERT; the curves are instead nearly superimposable. Small quantitative differences (SciBERT showing the largest shifts, BlueBERT the smallest) are better explained by differences in probe accuracy—a proxy for the sharpness of the uncertainty axis—rather than corpus content.

6 Limitations

Encoder generation. We measure representation shifts in encoder models only; behavioral impact on downstream NLP tasks (named entity recognition, relation extraction, information extraction) remains open. The $\approx 50\%$ flip rate confirms meaningful prediction changes at the probe level, but task-level evaluation is required to assess real-world impact.

Contrast set. The 400-sentence BioScope contrast set mixes genres (abstracts, full papers, clinical reports) and speculation types (epistemic hedging, conditionals, hypotheticals). A narrower annotation scheme focused specifically on clinical epistemic uncertainty might yield a sharper representation and higher steering fidelity.

Orthogonalization. Only sentence length and hedge-word count are controlled; specific modal constructions, passive voice, negation scope, and syntactic bigrams remain uncontrolled confounds. The claim that uncertainty encoding survives orthogonalization should be read as holding for these two confounds specifically.

Hedge metric. Our 27-term lexicon detects the presence of hedging markers but not their clinical appropriateness. A model generating repetitive hedging vocabulary outside coherent clinical content would receive a high score. Expert radiologist evaluation of output quality is an essential next step for any clinical safety claim.

Single prompt domain. Decoder experiments use radiology prompts exclusively. Steering behavior may differ for other clinical genres such as pathology reports, discharge summaries, or progress notes, where the distribution of uncertainty cues differs substantially.

Window size. The ± 2 window is not cross-validated across held-out model families. Results are empirically insensitive to ± 1 or ± 3 for all six models in this study, but the optimal window width may differ for models with sharper or flatter probe accuracy curves.

References

Heike Adel and Hinrich Schütze. 2015. Comparing approaches to the coreference resolution of clinical speculation and negation. In *Proceedings of the BioNLP 2015 Workshop*.

Shashank Agarwal and Hongfang Yu. 2010. Detecting hedging in biomedical text using a semantic-based approach. In *Proceedings of the BioNLP 2010 Workshop*.

Emily Alsentzer, John R. Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. BioMedLM: A 2.7b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Jiayang Geng, Fei Cai, Yida Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

David M. Panicek and Hedvig Hricak. 2016. How the language of radiology reports can foster (or foil) clinical communication. *Journal of the American College of Radiology*, 13(12):1394–1395. Editorial.

Arjun Panickssery, Nick Gabrieli, Julia Bowman, Meg Tong, Evan Hubinger, and Alexander Mallen. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3363–3377.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Alex Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical hallucination test bed for large language models. *arXiv preprint arXiv:2307.15343*.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7222–7240.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Experimental Details

Hyperparameters. Generation: temperature 0.8, top-p 0.9, max new tokens 60, min new tokens 20, repetition penalty 1.2. Probe: $C = 0.1$, max_iter 2000, random_state 42. 5-fold stratified CV, seed 42. Sample sizes: 200+200 BioScope sentences for probe training; 80 balanced sentences (held out) for encoder steering; 20 seeds $\times$ 5 prompts $=$ 100 generations per decoder configuration.

Reproducibility. Fixed random seeds throughout. BioScope XML files enumerated alphabetically before sampling; contrast set is platform-independent. Code and results available at the repository linked in the first-page footnote.

B Implementation Details

Compute environment. Python 3.11.9, PyTorch 2.5.1+cu121, Transformers 4.49.0, Linux 6.1.0, NVIDIA GPU, CUDA 12.1. Float32 is the canonical dtype.

Dtype robustness. bfloat16 and float16 produce probe accuracies within 0.25pp of float32 on the same CUDA device. BioMedLM bfloat16 shows minor steering drift at $\alpha_{\text{frac}} \geq 0.20$; qualitative findings are unchanged.

Cross-platform note. CUDA and Apple MPS use different floating-point kernels and RNG sequences, producing numerically distinct hidden states even with identical seeds. Within a fixed CUDA platform, results are reproducible.

C Numerical Precision Robustness

Table 4 reports probe accuracy at the selected steering layer $\hat{\ell}$ under three floating-point precisions on the same CUDA device. All six models return identical accuracies across float32, float16, and bfloat16, confirming that numerical precision does not affect the qualitative or quantitative probing findings. Minor steering drift is observed for BioMedLM bfloat16 at $\alpha_{\text{frac}} \geq 0.20$ (hedge score within ± 0.02 of float32); all qualitative findings are unchanged. Float32 is the canonical dtype for all reported results.

D Hedge Marker Lexicon

may, might, could, would, should; suggest, suggests, suggestive; possible, possibly, possibility; probable, probably, probability; likely, unlikely;

Table 4: Probe accuracy at $\hat{\ell}$ across floating-point precisions (same CUDA device). All differences are zero across all models.

Model	$\hat{\ell}$	float32	float16	bfloat16
BioGPT	16	85.50%	85.50%	85.50%
BioMedLM	22	84.25%	84.25%	84.25%
BioBERT	10	86.50%	86.50%	86.50%
ClinicalBERT	9	82.75%	82.75%	82.75%
BlueBERT	8	80.75%	80.75%	80.75%
SciBERT	10	85.00%	85.00%	85.00%

appear, appears, appearing; seem, seems; suspect, suspected, suspicion; consistent with, cannot exclude, rule out.

E Formal Evaluation Metrics

Lexical diversity.

$$D(g) = \frac{|\text{unique words}(g)|}{|\text{words}(g)|} \tag{9}$$

Perplexity.

$$\text{PPL}(g) = \exp\left(-\frac{1}{|g|}\sum_{t=1}^{|g|}\log p_\theta(g_t \mid g_{<t})\right) \tag{10}$$

where p_θ is BioGPT evaluated without steering hooks.

Token validity.

$$V(g) = \frac{|\{t \in g : t \in \mathcal{W}\}|}{|\{t \in g : t \text{ alphabetic}\}|} \tag{11}$$

where $\mathcal{W}$ is the set of valid English words (checked against the NLTK words corpus).

Length.

$$L(g) = |g| \tag{12}$$

Encoder steering metrics.

$$\Delta_{\text{proj}} = (\mathbf{h}^{\text{mod}} - \mathbf{h}^{\text{orig}}) \cdot \hat{\mathbf{v}}_{\text{probe}} \tag{13}$$

$$\Delta_{\text{prob}} = p(y{=}1 \mid \mathbf{h}^{\text{mod}}) - p(y{=}1 \mid \mathbf{h}^{\text{orig}}) \tag{14}$$

$$\text{flip}\% = \frac{|\{s : \hat{y}^{\text{mod}} \neq \hat{y}^{\text{orig}}\}|}{|S|} \tag{15}$$

$$\Delta_{\cos} = \cos(\mathbf{h}^{\text{mod}}, \bar{\mathbf{h}}_u) - \cos(\mathbf{h}^{\text{orig}}, \bar{\mathbf{h}}_u) \tag{16}$$

$$f_{\text{axis}} = \frac{|\Delta\mathbf{h} \cdot \hat{\mathbf{v}}_{\text{probe}}|}{\|\Delta\mathbf{h}\|} \tag{17}$$

where $\bar{\mathbf{h}}_u$ is the mean hidden state of uncertain training sentences, $\hat{y}$ denotes the probe's predicted class, and $\Delta\mathbf{h} = \mathbf{h}^{\text{mod}} - \mathbf{h}^{\text{orig}}$.

F Full Probing Results

Tables 5, 6, and 7 list layer-by-layer 5-fold CV probe accuracies for all six models. *=2/3 anchor; ^=window-selected. Figures 6–7 show bar-chart views of the decoder sweeps.

Table 5: BioGPT probe accuracy, all 25 layers. $\ell^* = 16$, $\hat{\ell} = 16$.

Layer	CV Acc	Std	95% CI
0	82.00%	4.38%	[78.16%, 85.84%]
1	84.50%	4.47%	[80.58%, 88.42%]
2	84.50%	3.26%	[81.64%, 87.36%]
3	84.25%	3.01%	[81.61%, 86.89%]
4	84.75%	2.24%	[82.79%, 86.71%]
5	86.25%	1.98%	[84.52%, 87.98%]
6	85.50%	2.09%	[83.67%, 87.33%]
7	84.25%	3.38%	[81.29%, 87.21%]
8	84.00%	2.40%	[81.89%, 86.11%]
9	85.50%	2.59%	[83.23%, 87.77%]
10	84.75%	2.40%	[82.64%, 86.86%]
11	84.50%	2.44%	[82.36%, 86.64%]
12	86.00%	1.85%	[84.37%, 87.63%]
13	84.75%	1.63%	[83.32%, 86.18%]
14	84.75%	2.05%	[82.95%, 86.55%]
15	85.00%	1.98%	[83.27%, 86.73%]
16*^	**85.50%**	**2.44%**	**[83.36%, 87.64%]**
17	84.75%	2.24%	[82.79%, 86.71%]
18	85.25%	2.85%	[82.75%, 87.75%]
19	85.50%	2.44%	[83.36%, 87.64%]
20	84.75%	3.11%	[82.02%, 87.48%]
21	84.25%	2.88%	[81.73%, 86.77%]
22	82.00%	2.27%	[80.01%, 83.99%]
23	81.50%	2.24%	[79.54%, 83.46%]
24	82.50%	2.50%	[80.31%, 84.69%]

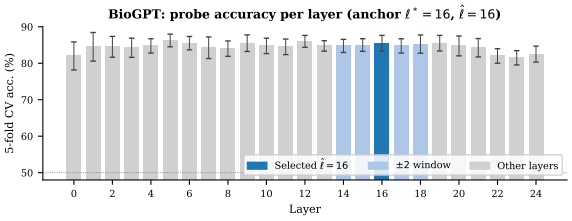


Figure 6: BioGPT probe accuracy per layer. Blue = selected L16; light blue = ±2 window; grey = other layers.

G Detailed Steering Results

Figure 10 shows the steering geometry: ortho direction on-axis fraction (0.97–0.98, constant across α) and probe direction $\Delta_{\cos}$ (sign reversal above $\alpha \approx 0.10$).

Table 6: BioMedLM probe accuracy, all 33 layers. $\ell^* = 21$, $\hat{\ell} = 22$.

Layer	CV Acc	Std	95% CI
0	55.25%	3.69%	[52.02%, 58.48%]
1	81.25%	2.65%	[78.93%, 83.57%]
2	82.50%	3.19%	[79.71%, 85.29%]
3	84.25%	3.38%	[81.29%, 87.21%]
4	83.00%	2.09%	[81.17%, 84.83%]
5	84.50%	3.14%	[81.75%, 87.25%]
6	84.75%	1.85%	[83.12%, 86.38%]
7	85.25%	3.47%	[82.21%, 88.29%]
8	84.75%	2.85%	[82.25%, 87.25%]
9	83.75%	2.93%	[81.18%, 86.32%]
10	83.25%	2.44%	[81.11%, 85.39%]
11	83.75%	1.98%	[82.02%, 85.48%]
12	84.50%	2.27%	[82.51%, 86.49%]
13	85.75%	2.09%	[83.92%, 87.58%]
14	85.50%	2.44%	[83.36%, 87.64%]
15	86.00%	1.05%	[85.08%, 86.92%]
16	86.00%	1.37%	[84.80%, 87.20%]
17	84.25%	1.90%	[82.59%, 85.91%]
18	83.75%	2.17%	[81.85%, 85.65%]
19	84.00%	1.85%	[82.37%, 85.63%]
20	83.75%	1.53%	[82.41%, 85.09%]
21*	84.00%	2.40%	[81.89%, 86.11%]
22$^\wedge$	**84.25%**	**1.90%**	**[82.59%, 85.91%]**
23	83.50%	2.98%	[80.88%, 86.12%]
24	83.50%	2.24%	[81.54%, 85.46%]
25	83.50%	3.47%	[80.46%, 86.54%]
26	83.00%	3.14%	[80.25%, 85.75%]
27	83.25%	3.38%	[80.29%, 86.21%]
28	82.75%	3.79%	[79.43%, 86.07%]
29	82.50%	4.15%	[78.87%, 86.13%]
30	82.25%	4.54%	[78.27%, 86.23%]
31	80.50%	5.05%	[76.08%, 84.92%]
32	81.00%	4.71%	[76.87%, 85.13%]

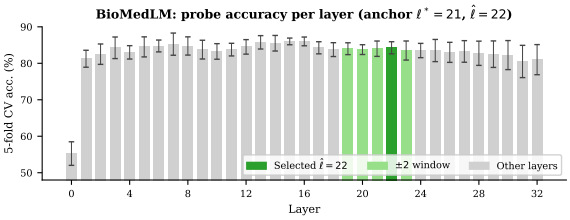


Figure 7: BioMedLM probe accuracy per layer. Green = selected L22; light green = ±2 window.

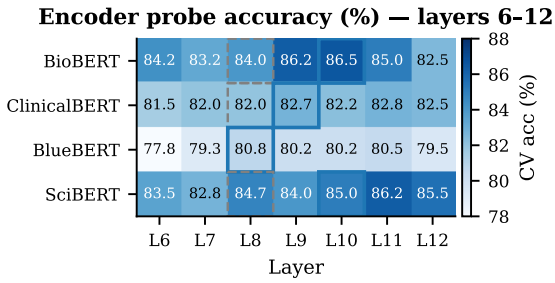


Figure 8: Encoder probe accuracy heatmap for layers 6–12. Blue boxes mark selected $\hat{\ell}$; dashed boxes mark the 2/3 anchor L8.

Table 7: Encoder probe accuracy (%) across all 13 layers. Anchor $\ell^* = 8$; *=anchor; $^\wedge$=window-selected; $^{*\wedge}$=anchor and window-selected coincide.

L	BioBERT	ClinBERT	BlueBERT	SciBERT
0	76.0±2.2	77.8±3.2	77.0±2.4	80.5±4.3
1	82.2±3.8	80.5±2.9	79.5±2.4	81.5±2.4
2	80.2±2.1	81.8±3.6	79.2±3.8	83.5±3.5
3	82.5±3.3	83.2±5.1	80.5±3.6	83.2±5.0
4	83.2±4.2	83.0±4.1	79.5±5.3	83.2±4.7
5	85.0±5.4	81.8±4.1	78.5±4.2	84.0±5.0
6	84.2±4.5	81.5±2.7	77.8±4.2	83.5±2.9
7	83.2±4.5	82.0±2.1	79.2±5.4	82.8±2.1
8*	84.0±4.5	82.0*±3.6	**80.8**$^{*\wedge}$±5.4	84.8*±4.5
9	86.2±3.6	**82.8**$^\wedge$±3.8	80.2±2.2	84.0±3.9
10$^\wedge$	**86.5**$^\wedge$±2.4	82.2±2.2	80.2±2.6	**85.0**$^\wedge$±3.4
11	85.0±4.0	82.8±2.4	80.5±1.4	86.2±3.6
12	82.5±2.7	82.5±2.7	79.5±1.7	85.5±3.4

Acc in %; ± denotes std over 5-fold CV.

Table 8: BioGPT decoder steering, layer 16, calibrated norms ($\|\mathbf{h}_{16}\| \approx 205$). Bold = peak hedge.

α	Hedge	Lex.div.	Tok.val.	PPL
Probe direction				
0.000	0.35	0.95	0.86	17.0
0.025	0.45	0.94	0.86	17.7
0.050	0.55	0.95	0.87	16.6
0.075	0.73	0.95	0.89	18.0
0.100	0.78	0.95	0.88	19.7
0.125	**0.90**	0.94	0.89	20.3
0.150	0.84	0.95	0.89	21.1
0.200	0.80	0.94	0.91	32.9
0.250	0.88	0.94	0.90	43.7
Ortho direction				
0.000	0.35	0.95	0.86	17.0
0.025	0.47	0.95	0.87	17.4
0.050	0.52	0.95	0.88	17.4
0.075	0.70	0.95	0.88	18.3
0.100	0.81	0.95	0.88	19.6
0.125	**0.97**	0.94	0.89	21.7
0.150	0.88	0.95	0.89	23.5
0.200	0.94	0.94	0.89	32.0
0.250	0.89	0.91	0.91	44.8

Table 9: BioMedLM decoder steering, layer 22, calibrated norms ($\|\mathbf{h}_{22}\| \approx 228$). Bold = peak hedge.

α	Hedge	Lex.div.	Tok.val.	PPL
Probe direction				
0.000	0.32	0.82	0.90	12.7
0.050	0.37	0.84	0.90	14.0
0.100	0.42	0.83	0.90	15.8
0.150	0.45	0.82	0.90	19.0
0.200	0.52	0.78	0.92	22.6
0.250	**0.61**	0.75	0.93	26.0
Ortho direction				
0.000	0.32	0.82	0.90	12.7
0.050	0.31	0.84	0.90	14.1
0.100	0.50	0.82	0.89	15.1
0.150	0.41	0.81	0.91	17.8
0.200	0.53	0.80	0.91	21.2
0.250	**0.44**	0.76	0.94	25.3

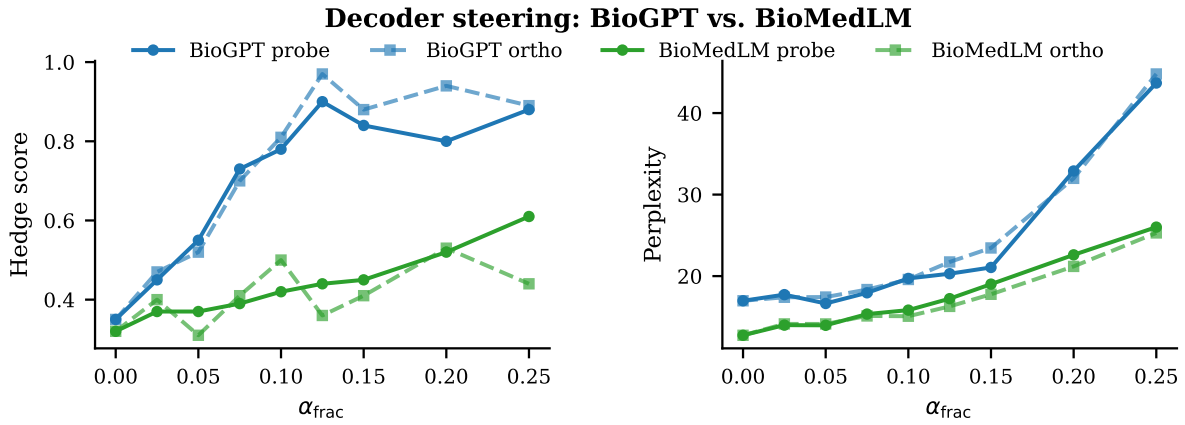


Figure 9: Side-by-side decoder steering comparison: BioGPT (L16, $\|\mathbf{h}\| \approx 205$, $d = 1024$) vs. BioMedLM (L22, $\|\mathbf{h}\| \approx 228$, $d = 2560$), probe and ortho directions. Left: hedge score. Right: perplexity. BioGPT exhibits a non-monotonic peak and steeper perplexity rise; BioMedLM shows a monotonic, collapse-free profile.

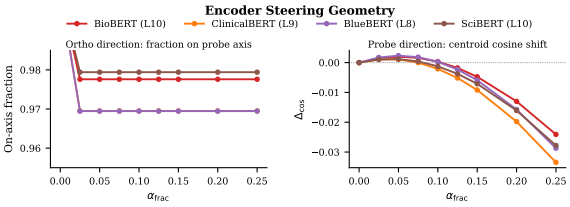


Figure 10: Left: ortho direction on-axis fraction (constant, 0.97–0.98 depending on model). Right: probe direction $\Delta_{\cos}$ showing sign reversal above $\alpha \approx 0.10$.

Table 10: Encoder steering (probe direction) at selected layers, averaged over 80 test sentences. On-axis = 1.000 for all entries (probe direction by construction).

Model	α	Δ_{proj}	Δ_{prob}	flip%	$\Delta_{\cos}$
BioBERT L10	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+0.99	+0.336	41.2%	+0.002
	0.10	+1.98	+0.507	52.5%	+0.000
	0.15	+2.97	+0.532	52.5%	−0.005
	0.25	+4.96	+0.534	52.5%	−0.024
ClinBERT L9	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+1.12	+0.381	45.0%	+0.001
	0.10	+2.24	+0.507	52.5%	−0.002
	0.15	+3.36	+0.519	52.5%	−0.009
	0.25	+5.60	+0.520	52.5%	−0.034
BlueBERT L8	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+0.77	+0.298	40.0%	+0.002
	0.10	+1.55	+0.472	51.2%	+0.000
	0.15	+2.32	+0.517	51.2%	−0.006
	0.25	+3.87	+0.525	51.2%	−0.029
SciBERT L10	0.00	+0.00	+0.000	0.0%	+0.000
	0.05	+1.41	+0.400	47.5%	+0.001
	0.10	+2.83	+0.528	53.8%	−0.001
	0.15	+4.24	+0.535	53.8%	−0.007
	0.25	+7.07	+0.535	53.8%	−0.028

Relations of Linguistic Features to Guideline-Based Rankings of Medical Answers are Nontrivial

Davis Bartels, Brandon C. Colelough, and Dina Demner-Fushman
National Library of Medicine, NIH
firstname.lastname@nih.gov

Abstract

We present an exploratory analysis of how simple linguistic features relate to annotator rankings of machine-generated medical answers under task-specific evaluation guidelines. We examine eight interpretable features of the answer text: length in words, average words per sentence, percentage of polysyllabic words, medical named entity density, perplexity, coherence, lexical diversity, and dependency distance. Our preliminary results show how the role of certain features depends on the ranking guidelines. When the guidelines explicitly instruct annotators to prefer shorter answers if informativeness is equal, annotators tend to be broadly consistent with the stated criteria, though with substantial annotator-level variation. In rankings where brevity is not part of the stated criteria, length still shows a strong, but less consistent association, suggesting that answer length may influence rankings even outside explicit guideline instructions. Other features show weaker and more variable associations. These findings suggest that linguistic features may provide insight into how annotators apply ranking guidelines, but that these relationships are heterogeneous and not fully explained by simple linear trends. The results motivate further analysis of guideline adherence, annotator variation, and possible confounding between linguistic form and answer content in medical answer evaluation.

1 Introduction

Human judgments of answer quality not only reflects whether an answer is correct, but also how that answer is written. Prior work in community question answering has shown that linguistic properties such as answer length, relevance to the question, and readability can influence which answers users prefer or select as best (Beygelzimer et al., 2015; Banjar et al., 2024). The influence of writing style on task-specific quality judgments is especially important in the medical domain, where readers must evaluate information that is often technical and difficult to process. Prior studies have shown that standard readability formulas do not fully capture the difficulty of medical text, and that features beyond simple word and sentence length may be needed to better characterize how understandable a document is (Wu et al., 2013; Zheng and Yu, 2018). Relatedly, Roberts and Demner-Fushman (2016) demonstrate the value of analyzing medical questions through multiple linguistic levels, comparing consumer and professional health questions using lexical, readability, language-model, semantic-type, and question-decomposition features. The findings from Roberts *et al.* demonstrate that medical text differs not only in surface readability, but also in terminology, background information, question structure, and semantic focus. Motivated by evidence that medical text varies across linguistic, semantic, and readability dimensions, we examine how a small set of interpretable linguistic features is associated with annotator rankings under task-specific guidelines. Our exploratory analysis is intended to establish whether associations are consistent with the guidelines, vary across annotators, or point to additional factors requiring future study.

2 Data

We use the ClinIQLink dataset, which comprises textbook-grounded medical questions each paired with eight machine-generated answers from a diverse set of open-source and internal language model systems (Colelough et al., 2025). Specifically, we draw from Task 2 of the dataset, a medical student ranking study that restricts evaluation to four open-ended question types, namely short answer, short inverse, multi-hop, and multi-hop inverse. Students from accredited medical schools ranked each set of eight answers according to a provided set of task-specific guidelines. We focus on the short answer and short inverse question types

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1080–1088
July 3-4, 2026 ©2026 Association for Computational Linguistics

Table 1: Example question and selected ranked model-generated answers.

Question: *What type of cancers do gene-based tests serve as the standard of care?*		
Rank	**Judgment**	**Answer**
1	Good	Gene-based tests are standard of care in certain breast (BRCA1, BRCA2), ovarian (BRCA1, BRCA2), colon (MLH1, MSH2, MSH6, PMS2, EPCAM), and lung (EGFR, ALK, ROS1, BRAF, KRAS) cancers, as well as melanoma (BRAF) and thyroid (BRAF, RET) cancers. They help determine risk, diagnosis, prognosis, and targeted therapy options.
2	Good	Gene-based tests serve as the standard of care for certain types of cancers, including breast cancer (BRCA1/BRCA2 mutations), colorectal cancer (MLH1, MSH2, MSH6, PMS2 mutations), ovarian cancer (BRCA1/BRCA2 mutations), prostate cancer (BRCA1/BRCA2, HOXB13 mutations), and lung cancer (EGFR, ALK, ROS1 mutations). These tests help in identifying hereditary cancer syndromes and guiding treatment decisions.

in our preliminary results, as these two types offer the clearest contrast in annotation guidelines. For short answer questions, annotators were instructed to rank the shorter of two equally informative answers higher, whereas for short inverse questions, annotators were instructed to rank answers solely by accuracy, strength of argument, and grounding. In addition to ranking, annotators were asked to assign a judgment to each answer, choosing from "Good", "Ok", or "Bad". "Good" indicates that the answer is complete and correct, "Ok" indicates the answer is acceptable with slight discrepancy, and "Bad" indicates that the answer is not factually accurate or does not address the question. To reduce the influence of overt factual incorrectness and focus on variation among acceptable answers, only answers judged "Good" were considered for this study. Examples of ranking may be found in Table 1.

3 Methods

We define the below set of metrics for linguistic features to analyze in the answer texts. We aim to see how they associate with the annotators' rankings under the provided guidelines. The metrics were selected to sample a range of commonly studied linguistic features and reveal preliminary patterns, before expanding to a broader feature set in future work. The linguistic features we use to analyze the CLinIQLink dataset include:

Lexical Count We measure the length of the text in terms of the number of words.

Average Words Per Sentence We measure sentence length as the average number of words in a sentence.

Percentage Polysyllabic Words The percentage of polysyllabic words is frequently included as a component of readability metrics, defined as the percentage of words with three or more syllables. If a word exists in CMUdict (Bartlett et al., 2009), we count the number of vowel phonemes in the word. If it does not, we approximate syllable count using Hunspell hyphenation dictionaries.

Medical Entity Density Using MetaMapLite (Demner-Fushman et al., 2017), we record the number of unique spans recognized as a medical named entity. We compute density by normalizing by the number of words in the text.

Dependency Distance Dependency distance, described by Oya (2011), is used as a measure of syntactic complexity, where a greater distance represents greater complexity. We average over the dependency distance for each sentence.

Coherence We measure the coherence between two sentences as the cosine similarity between their embedding vectors. We use `all-MiniLM-L6-v2` [1] to produce embedding vectors. We average over the coherence for each pair of consecutive sentences.

MATTR To capture lexical diversity, we measure the Moving-Average Type-Token Ratio (MATTR) of the text (Covington and McFall, 2010). MATTR is the average of the ratio of unique words to total words in a specified window size, for each window in the text. We use a window size of 50.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Perplexity Perplexity is the exponentiated average negative log-likelihood of a sequence. We use GPT-2 to compute the log-likelihoods for each token.

3.1 Additive Utility Paired Comparison Model

To measure the relationships between each metric and the orderings, the ranked lists are converted into pairwise data. We model each answer i's latent utility, η_i, as an additive function of its M metrics of linguistic features. Specifically, x_{im} denotes the value of feature m for answer i, and $f_m(\cdot)$ is a feature-specific spline function that captures the potentially nonlinear contribution of feature m to the answer's overall utility, as shown in Equation 1:

$$\eta_i = \sum_{m=1}^{M} f_m(x_{im}) \tag{1}$$

We then model the probability that answer i is ranked over answer j, denoted $P(i \succ j)$, as a logistic function of the difference between their utilities, $\eta_i - \eta_j$, as shown in Equation 2:

$$P(i \succ j) = \text{logit}^{-1}(\eta_i - \eta_j) \tag{2}$$

Thus, answers with higher estimated utility are more likely to be preferred in pairwise comparisons.

3.2 Normalization and Cross-Validation

Metrics are first normalized and missing data, such as MATTR for texts shorter than the window size or coherence for texts with only one sentence, are imputed with the average value for that feature after normalization. A missingness indicator is introduced as a covariate for each feature with imputed values. We use 5-fold cross validation to reduce variance in our performance metrics and split folds by question to prevent leakage. Annotators with fewer than 20 ranked lists for a particular question type were excluded for that question type.

3.3 Annotator Summary Metrics

We summarize the models across annotators through a variety of metrics. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators, giving an overall direction of the curve for that feature. We count the percentage of annotators that exhibit a positive and negative direction for a given feature with a threshold of 0.05, such that a value x is not counted if $-0.05 \leq x \leq 0.05$. Utility range is $\max f(x) - \min f(x)$ and nonlinearity is measured as the Root Mean Square Error between the estimated smooth curve and its best-fitting straight line, normalized by the range of the smooth curve. To analyze annotator agreement, we use curve correlation and curve variability. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean of the standard deviations of the dispersion of annotator-specific smooth curves across the feature range.

3.4 Feature Ablation

To assess the importance of each feature to the model, we conduct a leave-one-out ablation study. We fit the model again, with one feature removed, and measure the change in held-out negative log-likelihood. Positive values indicate that removing the feature worsened predictive performance.

4 Results

4.1 Short Answer Feature Relationships

Table 2 summarizes the feature relationship metrics across annotators for the short answer questions. Lexical count shows the strongest directional indication with a median direction of -0.238, and also the largest utility range at 1.168, though its IQR is large at $[-0.747, -0.081]$, reflecting substantial variation across annotators. A large percentage of annotators demonstrate a negative direction for both lexical count (76.5%) and dependency distance (70.6%), while perplexity has the largest percentage of annotators below the threshold at 52.9%. None of the remaining features show strong directional indication. Nonlinearity is broadly similar across all features.

4.2 Short Answer Curve Agreement and Ablation Results

Table 3 summarizes curve agreement and ablation results for the short answer questions. Lexical count and dependency distance mean have the greatest curve correlations at 0.824 and 0.822 respectively, though only lexical count exhibits high curve variability at 0.439, suggesting that even among annotators for whom length is a strong predictor, the shape of the relationship or magnitude of the effect size differs considerably. All features show very low magnitude changes in negative log-likelihood across the ablation study.

Table 2: Annotator-level feature relationship summaries for short answer questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.238 [-0.747, -0.081]	17.6	76.5	1.168 [0.834, 1.795]	0.084 [0.017, 0.131]
Pct. Polysyllabic Words	-0.005 [-0.112, 0.108]	41.2	35.3	0.641 [0.396, 0.875]	0.129 [0.098, 0.216]
Avg. Words Per Sentence	-0.082 [-0.175, 0.080]	41.2	52.9	0.637 [0.393, 0.971]	0.106 [0.067, 0.190]
Medical Entity Density	-0.059 [-0.127, 0.033]	23.5	52.9	0.570 [0.321, 0.678]	0.122 [0.068, 0.182]
MATTR	-0.058 [-0.249, 0.059]	35.3	52.9	0.496 [0.174, 0.618]	0.140 [0.109, 0.208]
Dependency Distance Mean	-0.090 [-0.158, 0.005]	17.6	70.6	0.486 [0.326, 0.642]	0.109 [0.072, 0.181]
Perplexity	0.000 [-0.072, 0.019]	11.8	35.3	0.413 [0.272, 0.665]	0.043 [0.020, 0.135]
Coherence	0.056 [-0.113, 0.099]	52.9	35.3	0.359 [0.237, 0.534]	0.162 [0.130, 0.180]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 3: Annotator-level curve agreement and ablation summaries for short answer questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.824 [-0.539, 0.990]	0.439	-0.002 [-0.017, 0.007]
Pct. Polysyllabic Words	-0.125 [-0.932, 0.900]	0.131	-0.001 [-0.004, 0.004]
Avg. Words Per Sentence	0.091 [-0.906, 0.973]	0.108	0.000 [-0.004, 0.003]
Medical Entity Density	0.421 [-0.940, 0.959]	0.085	-0.001 [-0.004, 0.003]
MATTR	-0.116 [-0.925, 0.930]	0.159	0.006 [-0.016, 0.023]
Dependency Distance Mean	0.822 [-0.955, 0.992]	0.067	-0.004 [-0.007, -0.001]
Perplexity	0.650 [-0.996, 0.999]	0.091	-0.000 [-0.004, 0.007]
Coherence	-0.196 [-0.939, 0.945]	0.127	-0.006 [-0.008, 0.021]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

4.3 Short Inverse Feature Relationships

Table 4 summarizes the feature relationship metrics across annotators for the short inverse questions. Lexical count again shows the strongest directional indication with a median direction of -0.301 and the largest utility range at 1.932, though the IQR is considerably wider at $[-1.126, 0.491]$ than for the short answer questions, reflecting a more even split between annotators favoring shorter and longer answers. Average words per sentence shows the next strongest negative direction at -0.118, while medical entity density is the only feature with a notably positive direction at 0.101, with 60.7% of annotators exhibiting a positive association. Nonlinearity remains broadly similar across all features.

4.4 Short Inverse Curve Agreement and Ablation Results

Table 5 summarizes curve agreement and ablation results for the short inverse questions. Lexical count and average words per sentence have the greatest curve correlations at 0.571 and 0.489 respectively, and lexical count again exhibits the highest curve variability at 0.496. Lexical count is the only feature to show a meaningful positive change in negative log-likelihood upon removal, with a median ablation ΔNLL of 0.019 $[0.003, 0.050]$.

5 Discussion

5.1 Length and Annotator Guidelines

For the short answer questions, lexical count shows a very strong negative direction, meaning annotators, on average, ranked shorter text over longer text. This trend is closely aligned with the guidelines asking annotators to rank shorter answers higher for the short answer question type. However, the large IQR reveals variation in degree of direction, meaning that some annotators may not have strictly adhered to the brevity guideline, at least for some questions. We also see a strong negative direction for short inverse questions, despite annotators not being asked to consider length in their rankings. The IQR is even larger, showing many annotators tended to rank longer text over shorter text, despite the overall trend. For the short inverse questions, the IQR points to annotators having a strong, yet varied, association between length and perceived accuracy, strength of argument, and grounding. It is worth noting that annotators could have been biased by the instruction to prefer shorter answers on the short answer questions.

5.2 Feature Associations

While lexical count is the only feature with significant direction for the short answer question, the short inverse answer rankings show modest asso-

Table 4: Annotator-level feature relationship summaries for short inverse questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.301 [-1.126, 0.491]	39.3	60.7	1.932 [1.109, 3.613]	0.075 [0.044, 0.116]
Perplexity	-0.017 [-0.169, 0.042]	25.0	42.9	0.684 [0.356, 0.832]	0.089 [0.055, 0.153]
Avg. Words Per Sentence	-0.118 [-0.189, 0.078]	28.6	60.7	0.661 [0.465, 0.875]	0.099 [0.057, 0.170]
MATTR	-0.032 [-0.230, 0.152]	35.7	50.0	0.631 [0.395, 0.870]	0.126 [0.101, 0.181]
Pct. Polysyllabic Words	-0.038 [-0.356, 0.086]	32.1	50.0	0.613 [0.347, 0.868]	0.134 [0.097, 0.195]
Medical Entity Density	0.101 [-0.083, 0.213]	60.7	32.1	0.501 [0.352, 0.739]	0.139 [0.084, 0.197]
Coherence	-0.032 [-0.127, 0.156]	35.7	46.4	0.371 [0.339, 0.488]	0.166 [0.139, 0.198]
Dependency Distance Mean	-0.051 [-0.144, 0.044]	21.4	50.0	0.304 [0.238, 0.494]	0.163 [0.092, 0.212]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 5: Annotator-level curve agreement and ablation summaries for short inverse questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.571 [-0.975, 0.985]	0.496	0.019 [0.003, 0.050]
Perplexity	0.339 [-0.916, 0.940]	0.092	-0.001 [-0.004, 0.001]
Avg. Words Per Sentence	0.489 [-0.921, 0.974]	0.122	-0.001 [-0.005, 0.003]
MATTR	-0.081 [-0.938, 0.940]	0.175	-0.002 [-0.007, 0.002]
Pct. Polysyllabic Words	-0.054 [-0.918, 0.895]	0.145	0.000 [-0.005, 0.003]
Medical Entity Density	0.160 [-0.934, 0.945]	0.140	-0.003 [-0.005, 0.001]
Coherence	-0.156 [-0.904, 0.897]	0.107	-0.002 [-0.008, 0.002]
Dependency Distance Mean	-0.055 [-0.840, 0.894]	0.090	-0.005 [-0.012, -0.002]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

ciation with shorter sentences and higher medical entity density. It is to be expected that we may see stronger relationships with more features when one specific feature is not singled out in the guidelines. Both the positive and negative directions claim a significant portion of annotators for the short inverse questions, which signal variation in the associations across annotators. The significant percentage of annotators with negative directions for dependency distance, despite a low direction value for the short answer questions, could mean agreement, but weak association, or a nonlinear relationship not captured by the direction metric.

5.3 Nonlinearity and Annotator Variation

All features, for both question types, show some degree of nonlinearity, meaning some may have complex relationships. There may be, for example, a parabolic relationship between a specific feature and an annotator's rankings that would not be observable in the direction metric. For the short answer questions, lexical count has a high curve correlation, but also high variability. This again suggests even across annotators for whom length is a strong predictor of ranking, the relationships are very different. This variance in relationship implies that the brevity guideline was not applied uniformly across annotators.

5.4 Ablation and Feature Concurvity

The leave-one-out ablation study describes only lexical count as having a significant positive difference in negative log-likelihood and only for the short inverse question type. Though some features likely truly have little impact on the model, the negligible deltas of others could be explained by concurvity among features. Despite normalizing for length or selecting metrics relatively robust to length, some other features likely carry some signal for length through specific behavior at extremely short lengths. The percentage of polysyllabic words, for example, must be either 0% or 100%, when lexical count is 1.

6 Conclusion

This analysis of linguistic features in medical text and annotator rankings shows strong relationships for some annotators, as well as broader trends across the cohort. These findings highlight the nontrivial relationships between linguistic features and ranking judgments under task-specific guidelines. Future work includes assessing the nature of specific complex relationships, the inclusion of a broader set of features, and more detailed analysis.

Limitations

There are far more linguistic features that are frequently analyzed in both medical and open-domain texts than included in this study. The answer choices for each question were generated by the same eight systems and system identity was not ruled out as a potential confounder. Furthermore, because answers not judged "Good" were excluded from the study, it may be that some trends are better explained by the factuality of the answer, rather than the features analyzed. Questions were not analyzed by medical domain, which will be necessary to understand if the results depend on the topic of the text. Finally, because this study infers guideline application from associations between linguistic features and rankings, it cannot determine why annotators deviated from guideline-consistent patterns. Such deviations may reflect misunderstanding, disagreement with the guideline, attention to unmeasured answer qualities, or residual differences in correctness or informativeness among answers judged "Good."

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Ameen Banjar, Awais Shaheen, Tehmina Amjad, Riad Alharbey, and Ali Daud. 2024. Users' satisfaction based ranking for yahoo answers. *Multimedia Tools and Applications*, 83(28):71265–71284.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316, Boulder, Colorado. Association for Computational Linguistics.

Alina Beygelzimer, Ruggiero Cavallo, and Joel Tetreault. 2015. On yahoo answers, long answers are best. In *Proceedings of the ICML 2015 Workshop on Crowdsourcing and Machine Learning (CrowdML)*.

Brandon Colelough, Davis Bartels, and Dina Demner-Fushman. 2025. Overview of the ClinIQLink 2025 shared task on medical question-answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 378–387, Vienna, Austria. Association for Computational Linguistics.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17:100 – 94.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Masanori Oya. 2011. Syntactic dependency distance as sentence complexity measure. *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *J. Am. Med. Inform. Assoc.*, 23(4):802–811.

Danny T. Y. Wu, David A. Hanauer, Qiaozhu Mei, Patricia M. Clark, Lawrence C. An, Jianbo Lei, Joshua Proulx, Qing Zeng-Treitler, and Kai Zheng. 2013. Applying multiple methods to assess the readability of a large corpus of medical documents. *Studies in Health Technology and Informatics*, 192:647–651.

Jiaping Zheng and Hong Yu. 2018. Assessing the readability of medical documents: A ranking approach. *JMIR Medical Informatics*, 6(1):e17.

A Appendix

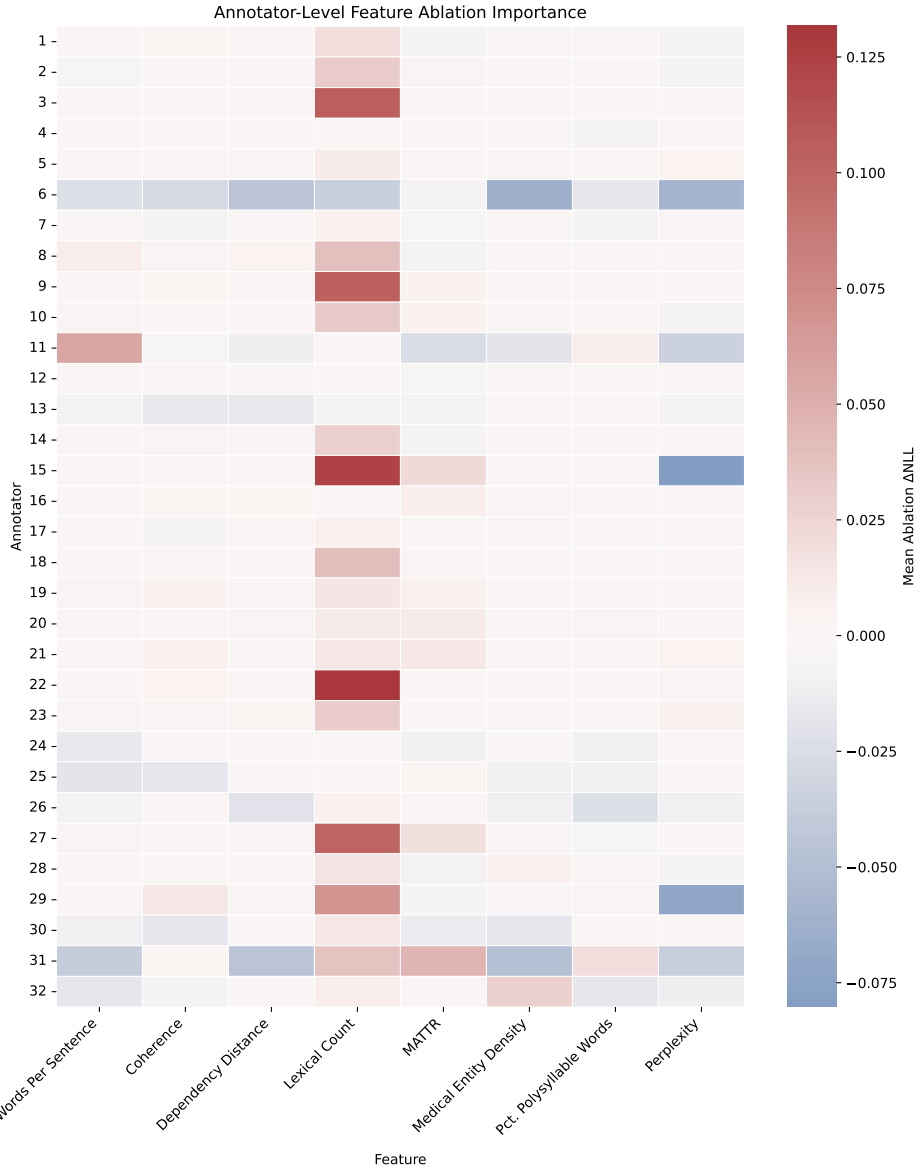


Figure 1: Heatmap of annotators and features. Color indicates the change in negative log-likelihood when each feature is removed.

Table 6: Annotator-level cross-validated model performance.

Annotator	Mean NLL	NLL SD	Pairwise acc.	Pairwise acc. SD	Kendall's τ	Kendall's τ SD
1	0.667	0.037	0.613	0.044	0.142	0.094
2	0.573	0.069	0.728	0.029	0.417	0.062
3	0.470	0.041	0.816	0.022	0.591	0.087
4	0.689	0.013	0.542	0.029	0.096	0.050
5	0.656	0.020	0.611	0.022	0.197	0.051
6	0.783	0.063	0.480	0.076	-0.012	0.113
7	0.698	0.011	0.520	0.033	0.037	0.031
8	0.524	0.075	0.760	0.078	0.483	0.145
9	0.416	0.030	0.818	0.020	0.589	0.053
10	0.621	0.058	0.662	0.036	0.321	0.071
11	0.692	0.072	0.577	0.046	0.129	0.087
12	0.698	0.005	0.508	0.016	0.016	0.035
13	0.719	0.013	0.479	0.036	-0.013	0.116
14	0.668	0.019	0.576	0.021	0.173	0.057
15	0.471	0.177	0.838	0.025	0.609	0.054
16	0.621	0.034	0.656	0.028	0.325	0.089
17	0.615	0.032	0.662	0.039	0.309	0.063
18	0.607	0.017	0.674	0.017	0.368	0.035
19	0.572	0.011	0.715	0.020	0.396	0.066
20	0.666	0.018	0.586	0.020	0.123	0.110
21	0.585	0.051	0.696	0.036	0.383	0.077
22	0.420	0.017	0.812	0.026	0.620	0.076
23	0.649	0.019	0.600	0.032	0.153	0.043
24	0.689	0.036	0.551	0.037	0.098	0.077
25	0.667	0.035	0.604	0.022	0.195	0.053
26	0.719	0.050	0.548	0.051	0.094	0.143
27	0.494	0.041	0.784	0.025	0.571	0.066
28	0.647	0.034	0.636	0.036	0.214	0.103
29	0.573	0.023	0.727	0.027	0.473	0.074
30	0.691	0.018	0.560	0.053	0.138	0.138
31	0.706	0.042	0.551	0.065	0.202	0.162
32	0.676	0.010	0.596	0.026	0.165	0.119

Note. Mean NLL, pairwise accuracy, and Kendall's τ are weighted cross-validated estimates. SD columns are fold-level standard deviations.

Table 7: Annotator-level feature relationship summaries for multi-hop questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.566 [-0.898, 0.039]	24.000	68.000	1.333 [0.684, 2.596]	0.079 [0.058, 0.145]
Coherence	-0.009 [-0.260, 0.097]	36.000	48.000	0.695 [0.444, 0.918]	0.159 [0.118, 0.204]
Medical Entity Density	-0.093 [-0.175, 0.146]	44.000	52.000	0.535 [0.392, 0.792]	0.136 [0.077, 0.201]
Dependency Distance Mean	-0.061 [-0.202, 0.118]	32.000	56.000	0.524 [0.256, 0.750]	0.155 [0.106, 0.185]
MATTR	-0.006 [-0.088, 0.152]	44.000	36.000	0.507 [0.293, 0.792]	0.168 [0.149, 0.212]
Pct. Polysyllabic Words	-0.043 [-0.168, 0.175]	36.000	48.000	0.462 [0.287, 0.609]	0.141 [0.099, 0.190]
Perplexity	0.003 [-0.079, 0.151]	36.000	36.000	0.441 [0.213, 0.760]	0.105 [0.067, 0.194]
Avg. Words Per Sentence	-0.008 [-0.087, 0.101]	28.000	36.000	0.430 [0.297, 0.700]	0.133 [0.095, 0.195]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 8: Annotator-level curve agreement and ablation summaries for multi-hop questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.715 [-0.899, 0.965]	0.337	0.012 [-0.001, 0.040]
Coherence	-0.039 [-0.828, 0.788]	0.193	-0.001 [-0.004, 0.002]
Medical Entity Density	-0.026 [-0.922, 0.946]	0.117	-0.002 [-0.009, -0.000]
Dependency Distance Mean	0.070 [-0.910, 0.902]	0.129	-0.003 [-0.005, 0.002]
MATTR	-0.068 [-0.912, 0.891]	0.132	-0.001 [-0.003, 0.003]
Pct. Polysyllabic Words	-0.438 [-0.953, 0.941]	0.120	-0.003 [-0.007, 0.000]
Perplexity	-0.662 [-0.962, 0.945]	0.112	-0.003 [-0.008, 0.000]
Avg. Words Per Sentence	0.011 [-0.902, 0.903]	0.071	-0.002 [-0.006, -0.001]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

Table 9: Annotator-level feature relationship summaries for multi-hop inverse questions. Values in brackets indicate interquartile ranges.

Feature	Direction	% Pos.	% Neg.	Utility range	Nonlinearity
Lexical Count	-0.136 [-0.874, 0.092]	33.333	62.500	1.800 [0.908, 3.066]	0.062 [0.040, 0.110]
Dependency Distance Mean	-0.010 [-0.146, 0.107]	45.833	50.000	0.514 [0.304, 0.637]	0.136 [0.074, 0.206]
Medical Entity Density	0.079 [0.012, 0.202]	62.500	16.667	0.507 [0.379, 0.767]	0.170 [0.091, 0.226]
Avg. Words Per Sentence	-0.034 [-0.183, 0.047]	25.000	45.833	0.494 [0.318, 0.716]	0.129 [0.096, 0.212]
Coherence	-0.075 [-0.184, 0.056]	29.167	54.167	0.451 [0.301, 0.591]	0.166 [0.112, 0.197]
MATTR	-0.053 [-0.172, 0.045]	25.000	50.000	0.393 [0.260, 0.648]	0.150 [0.089, 0.212]
Pct. Polysyllabic Words	-0.018 [-0.189, 0.137]	41.667	50.000	0.340 [0.262, 0.582]	0.142 [0.120, 0.200]
Perplexity	-0.001 [-0.127, 0.067]	37.500	41.667	0.287 [0.211, 0.891]	0.121 [0.063, 0.196]

Note. Direction is the median utility contrast $f(x_{75}) - f(x_{25})$ across annotators. Utility range is $\max f(x) - \min f(x)$. Nonlinearity is the normalized deviation from the best-fitting linear approximation.

Table 10: Annotator-level curve agreement and ablation summaries for multi-hop inverse questions. Values in brackets indicate interquartile ranges.

Feature	Curve corr.	Curve variability	Ablation ΔNLL
Lexical Count	0.244 [-0.966, 0.975]	0.408	0.029 [0.000, 0.069]
Dependency Distance Mean	-0.274 [-0.934, 0.870]	0.112	-0.003 [-0.009, 0.002]
Medical Entity Density	0.428 [-0.610, 0.878]	0.127	-0.002 [-0.006, 0.001]
Avg. Words Per Sentence	0.097 [-0.850, 0.900]	0.105	-0.002 [-0.008, 0.002]
Coherence	0.167 [-0.893, 0.902]	0.107	-0.005 [-0.009, -0.000]
MATTR	-0.108 [-0.839, 0.891]	0.116	-0.004 [-0.007, 0.005]
Pct. Polysyllabic Words	-0.737 [-0.963, 0.956]	0.109	-0.003 [-0.007, 0.000]
Perplexity	0.020 [-0.966, 0.956]	0.076	-0.001 [-0.009, 0.005]

Note. Curve correlation is the median pairwise correlation between annotator-specific smooth curves. Curve variability is the mean standard deviation of annotator-specific smooth curves across the feature range. Ablation ΔNLL is the change in held-out negative log likelihood when the feature is removed; positive values indicate that removing the feature worsened predictive performance.

Overview of the MedGenVidQA 2026 Shared Task on Medical Generative Video Question Answering

Deepak Gupta[1], **Collin S. Campbell**[2], **Pedram Golnari**[3], **Dina Demner-Fushman**[1]
[1] National Library of Medicine, National Institutes of Health, MD, USA
[2] Rowan-Virtua School of Osteopathic Medicine, NJ, USA
[3] Case Western Reserve University, OH, USA
[1]{first.last}@nih.gov
[2]campbe183@rowan.edu
[3]pxg338@case.edu

Abstract

This paper presents an overview of the MedGenVidQA 2026 shared task on medical video question answering, collocated with the 25th BioNLP workshop at ACL 2026. The shared task addressed three related sub-tasks of the medical multimodal (textual and video) question answering: **(i)** multimodal retrieval tasks, **(ii)** multimodal answer generation with citations, and **(iii)** a visual answer localization task. The key theme of the stated task is to develop reliable multimodal question-answering systems for consumers and medical professionals by leveraging generative models. A total of eight teams participated in the shared task challenges and submitted a total of forty-three submissions across all tasks. We performed both automated and human assessments to evaluate the submissions. This paper describes the tasks, datasets, evaluation metrics, participation, and baseline systems for all three tasks. Additionally, we summarize the techniques and results of the evaluation of the various approaches explored by the participating teams. Finally, we discuss the key findings and implications for the development of multimodal medical question answering.

1 Introduction

Recent advances in foundation models have significantly improved their ability to process, comprehend, and integrate information across multiple modalities such as text, images, and audio. These models have achieved superior performance on tasks that require aligning the modalities to capture the overall dynamics of the input sources. In the medical domain, knowledge is often distributed across a broader range of modalities, including textual sources such as scientific literature and clinical notes, as well as medical imaging, procedural, and instructional videos. Textual information often provides key insights, while visual data offers fine-grained spatial and temporal details, which are critical for understanding medical questions that demand a step-by-step procedure in the answer. Despite advances in state-of-the-art multimodal models, effectively combining heterogeneous knowledge sources remains a challenge, especially in the medical domain, where information from a single modality may not be sufficient to provide a comprehensive answer.

In the literature, multimodal question-answering tasks have been studied by incorporating images and videos as additional modalities alongside language. These tasks require understanding information across multiple modalities and their interactions to answer questions correctly. Most existing studies (Pramanick et al., 2024; Zhang et al., 2023; Liang et al., 2024) focus on open-domain settings involving videos (Lei et al., 2018, 2019; Rawal et al.; Zhang et al., 2024), images (Chowdhury and Soni, 2025; Tanaka et al., 2023), charts (Masry et al., 2025; Wu et al., 2024), and tables (Wang et al., 2024b; Pal et al., 2023). In the medical domain, image-based question answering has advanced with the development of neural-based approaches (Li et al., 2023; Hartsock and Rasool, 2024). There has also been some work on medical video-based question answering, including the creation of datasets (Gupta et al., 2023; Gupta and Demner-Fushman, 2024) and methods (Gupta et al., 2024a) aimed at better understanding medical videos and answering instructional questions. In addition, prior works (Gupta et al., 2025a, 2024b) have explored generating answers to medical questions from textual sources, such as scientific literature, using retrieval-augmented generation-based approaches to produce answers along with supporting citations. However, existing work on video question answering and multimodal question answering, both in open-domain and medical settings, lacks several important aspects. First, there is limited focus on retrieving relevant multimodal information sources, particularly combining textual and

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1089–1100
July 3-4, 2026 ©2026 Association for Computational Linguistics

video data in the medical domain. Second, existing approaches do not fully address generating answers that jointly consider multiple modalities while also providing supporting evidence, which is critical for ensuring reliability and trustworthiness in medical applications. Third, there is limited work on answering professional-level medical questions from medical procedural or instructional videos.

To bridge these gaps, we introduced the MedGenVidQA 2026 shared task[1], which aims to explore and develop efficient algorithms for medical question answering, considering the multimodal information (mainly text and videos). In the first task (multimodal retrieval) of the MedGenVidQA 2026 shared task, participants were asked to develop systems that retrieve relevant video and PubMed articles from multimodal sources containing the answer to the medical query. The second task (multimodal answer generation) aims to generate an answer that includes attributions from multimodal sources for each answer sentence. Finally, the third task (visual answer localization) aims to effectively localize the visual answer to the given medical or health-related question in a professional medical video.

2 MedGenVidQA 2026 Task Descriptions

MedGenVidQA comprises three related subtasks on video question answering. The illustration of the tasks is provided in Figure 1, 2, and 3 and details are as follows:

2.1 Task A: Multimodal Retrieval (MMR)

Given a medical query and a collection of multimodal sources, including textual documents and videos, the task is to retrieve relevant sources that can support answering the query. This involves aligning the query to the textual and video sources, which can serve as the basis for the task of answer generation. There is no requirement to use all modalities; either text, video, or both can be used.

2.2 Task B: Multimodal Answer Generation (MAG)

Given a medical query and a collection of multimodal sources, including textual documents and videos, the task is to generate an answer that is grounded in retrieved evidence from these sources. The generated response must include explicit attributions, with each sentence supported by cited references drawn from the multimodal sources. Similar to Task A, the model can use one or more modalities as needed.

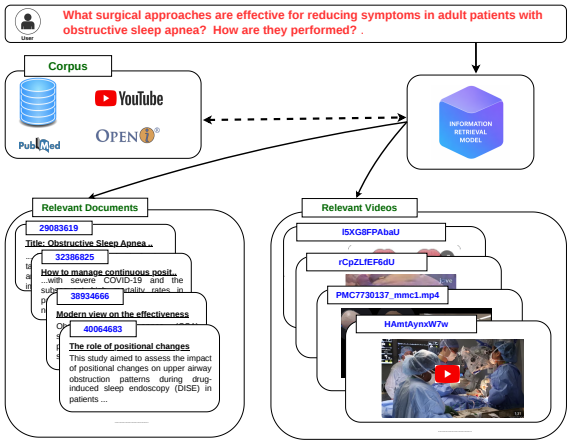


Figure 1: Example illustrating **MMR** task, where given the question, a retrieval model interacts with the corpus and is expected to produce the relevant video and documents as outputs.

2.3 Task C: Visual Answer Localization (VAL)

Given a medical query and a video, the task is to locate the temporal segments (start and end timestamps) in the video where the answer to the query is shown or the explanation is illustrated. This task assesses the model's ability to locate the visual answer in professional, procedure-based medical videos.

3 Data

3.1 Datasets

We developed textual and video corpora for the MedGenVidQA 2026 shared task. For the textual corpus, we used the latest annual baseline snapshot of MEDLINE/PubMed that covers the articles' abstracts and titles approximately through the end of 2025[2]. We provided a pre-processed set of $28,372,706$ PubMed abstracts.

In the video corpus, we focused on including both consumer and professional videos. For consumer-focused videos, we used the MedVidQA 2024 video corpus (Gupta and Demner-Fushman, 2024) of size $48,605$. This collection was developed by selecting videos from the "Personal Care and Style," "Health," and "Sports and Fitness" categories of the HowTo100M dataset (Miech et al.,

[1] https://medgenvidqa.github.io/

[2] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

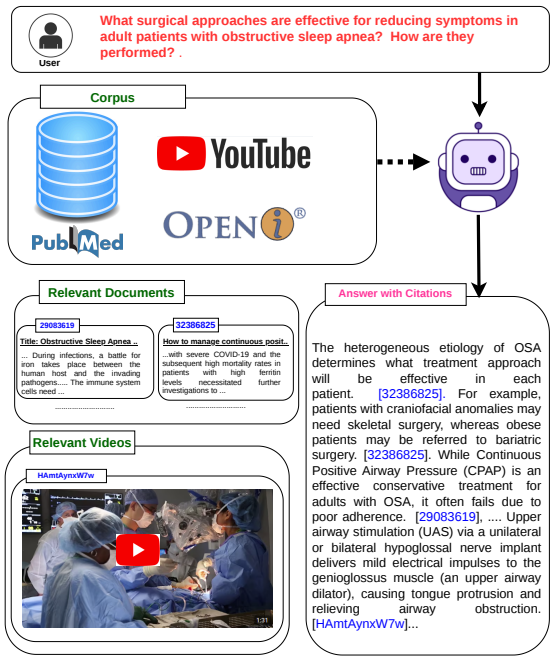


Figure 2: Illustration of the **MAG** task, where a model, given a question, can optionally use the corpus and is expected to produce an answer supported by sentence-level citations from relevant documents and videos.

2019), and further filtering instructional videos using the confidence scores from the instructional video classifiers developed in the MedVidQA corpus (Gupta et al., 2023). We augment the consumer-focused video corpus with medical professional videos from publicly available biomedical videos in the Open-i platform[3] associated with peer-reviewed publications in PubMed Central (PMC). These videos encompass a broad range of clinically relevant procedural and instructional content intended for scientific communication and education. We collected $10,462$ videos, each up to 3 minutes long. To facilitate multimodal learning, we also extracted transcripts for those videos using Whisper[4] with the `tiny` model version. The resulting video corpus, combining consumer and medical professional videos, contains $59,067$ videos. Participants were asked to use these collections (textual and video) to retrieve (Task A) and generate (Task B) answers to a given query.

[3] https://openi.nlm.nih.gov/

[4] https://github.com/SYSTRAN/faster-whisper

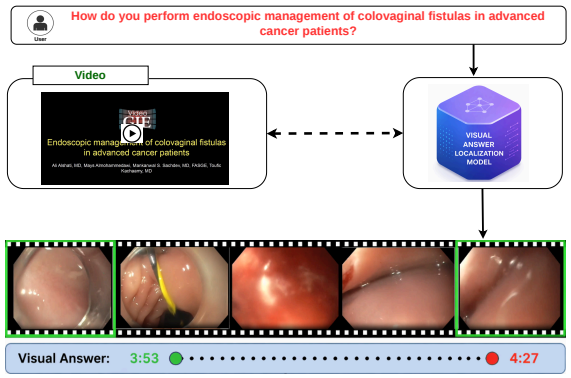


Figure 3: Illustration of the **VAL** task, where, given a question and a video, the model is expected to predict the start and end timestamps corresponding to the segment that serves as the visual answer to the question.

3.1.1 Training Set

For training and developing systems, we provided[5] the MedVidQA collections (Gupta et al., 2023) that consist of 3,010 human-annotated instructional questions and videos, along with visual answers from 900 health-related videos. The collection can support the training and development of all three tasks, as it includes medical questions, relevant videos (useful for Tasks A and B), and annotated temporal segments (useful for Task C) from those videos. We also provided the MedAESQA collection[6], comprising 8,427 human-annotated PubMed documents paired with answers to consumer health questions, where each answer sentence has up to three PubMed citations along with human-annotated relations with the PubMed documents in terms of *'support'*, *'contradict'*, *'neutral'*, and *'not relevant'*. The MedAESQA collection also provides human-annotated assessments of answer accuracy and relevancy ('*required*', '*unnecessary*', '*borderline*', and '*inappropriate*'), which can help develop systems for answer generation with citations (Task B). Together, MedVidQA and MedAESQA can be used to develop a multimodal question-answering system that retrieves relevant sources and uses them to generate answers with citations to those sources.

3.1.2 Test Set

Medical professionals on our team developed a new test set comprising 60 questions, ensuring a balance between consumer and medical professional needs. The questions were designed to ensure that both

[5] https://osf.io/pc594/files/osfstorage

[6] https://osf.io/ydbzq/files/osfstorage

textual and video references in the answers contribute to their informativeness and engagement for both audiences. For example, for the consumer health question *how do I take blood pressure?* Video sources can provide a clear demonstration of the procedure, including proper cuff placement and measurement technique, while textual sources can offer additional guidance on accuracy, recommended practices, and interpretation of readings. For medical professional questions such as *how is anesthesia intubation performed?*, video sources are essential for illustrating the procedural steps, whereas textual sources contribute important clinical details, including indications, safety considerations, and potential risks. We used the same set of 60 questions for Tasks A and B.

For Task C, a medical professional on our team sampled a set of 300 medical professional videos from Open-i, as discussed in Section 3.1. Each video was subjected to structured manual annotation to assess whether reasonable, clinically meaningful questions could be posed solely from visual content. If a video was found to contain a temporal segment that could be considered a visual answer to a medical question, the corresponding timestamps were recorded. This resulted in structured instances consisting of the question, the associated video, and the annotated temporal segment. In total, we constructed 80 question–video–answer instances for Task C.

4 Evaluation

4.1 Evaluation Metrics

4.1.1 MMR Evaluation

Participants were asked to retrieve relevant videos and abstract identifiers (PMIDs) (up to 10) for each question from both the video and textual corpora. To evaluate the relevant videos/text, we performed manual judgments using a pooling strategy on all system-retrieved videos (1,963) and PubMed documents (1,582). Using a pooling depth of $K = 3$, the final relevance pool consisted of 501 textual documents and 684 videos, with an average of 8.35 documents and 11.4 videos retrieved per topic, respectively. We instructed a total of four assessors with the following guidelines to assess the relevance of videos and documents:

Evaluating videos for relevance The videos are judged as being *Definitely Relevant*, *Possibly Relevant*, or *Not Relevant* to the given question. A video is considered *Definitely Relevant* if it contains a visual segment that constitutes a complete visual answer to the question. It is considered *Possibly Relevant* if it contains a visual segment that can be considered a partial/incomplete visual answer to the question. If the visual segments from the videos do not provide any visual answers to the question, the video can be marked as not relevant.

Evaluating PubMed documents for relevance Similar to the video relevance, the documents are also judged as *Definitely Relevant*, *Possibly Relevant*, or *Not Relevant* to the given question. *Definitely relevant* documents should directly answer the query and contain sufficient information on their own. *Possibly relevant* documents provide partial or related information but require additional sources to form a complete answer. *Not relevant* documents do not contribute useful information for answering the question.

We evaluated the performance of the video and document retrieval system in terms of Mean Average Precision (MAP), Recall@k, Precision@k, and nDCG metrics with $k = \{5, 10\}$. We follow the `trec_eval`[7] evaluation library to report the performance of participating systems.

4.1.2 MAG Evaluation

We used the BioACE (Gupta et al., 2026) evaluation framework[8] to assess the quality of system-generated answers. The evaluation focused on two aspects: completeness and correctness, measured using an automated approach and the ground-truth nuggets. Completeness measures how much of the required information (atomic biomedical facts, also called 'nuggets') is covered in the generated answer, whereas correctness assesses how well the generated content aligns with the relevant documents. In the nugget-based evaluation, nuggets are first extracted from the generated answer and then matched with the ground-truth nuggets using the relaxed matching scheme in BioACE. Precision (nugget-based correctness) and recall (nugget-based completeness) are then computed. These metrics are computed for each topic and then averaged over all questions in the test set to obtain the final scores. For citation evaluation, we used the BioACE framework. Specifically, we used LLaMA-3.3 to assess the relationship between each

[7] https://github.com/usnistgov/trec_eval

[8] https://github.com/deepaknlp/BioACE

answer sentence and its cited document or video sources. The model classifies each pair as *Supports*, *Contradicts*, or *Neutral*. Based on these labels, we computed citation coverage, citation support rate, and citation contradiction rate.

4.1.3 VAL Evaluation

Following Gupta et al. (2023); Gupta and Demner-Fushman (2022), we evaluated the performance of the VAL task using the following metrics:

Mean Intersection over Union (mIoU): For a given question q_i, IoU is computed as the ratio of the intersection area over the union area (Jaccard, 1912) between predicted and ground-truth temporal visual answer segments. It ranges from 0 to 1. A larger IoU means the predicted and ground-truth temporal visual answer segments match better, and IoU = 1.0 denotes an exact match. The mIoU is defined as the average temporal IoUs for all questions (N) in the test set. Formally,

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{i=N} IoU(q_i) \tag{1}$$

IoU @ μ is another metric used to evaluate the performance of the VAL system. It denotes the percentage of questions for which the predicted temporal segment has an IoU with the ground-truth segment greater than μ. Formally,

$$IoU@\mu = \frac{1}{N}\sum_{i=1}^{i=N} s(q_i, \mu), : \text{and} \tag{2}$$

$$s(q_i, \mu) = \begin{cases} 1, & \text{if } IoU(q_i) \geq \mu \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

We evaluated the submissions by considering $\mu = 0.3, 0.5, 0.7$. Since IoU@0.7 is the most restrictive among these settings, we used IoU@0.7 as the primary metric for ranking the submissions.

4.2 Baseline Systems

4.2.1 MMR Baselines

We developed baseline approaches using a BM25-based retrieval approach for both textual and video sources. For textual retrieval, we built an index over the PubMed corpus. For video retrieval, we constructed a separate index from the video transcriptions. Both indices were created using Pyserini[9] with default hyperparameters. Given a query, the system first retrieved top-100 relevant documents and videos independently from the respective indices using BM25, followed by a re-ranking step to refine the results using a pre-trained re-ranker model[10]. Building on this, we constructed two monomodal baselines, `Text-RR` and `Video-RR`, in which the top-10 documents after reranking are retained.

4.2.2 MAG Baselines

We adopted a RAG-based approach as developed in (Gupta et al., 2025b), which first fine-tuned a LLaMA2-7B (Touvron et al., 2023) model to generate answers with citations. The fine-tuning datasets were created from the PLABA (Attal et al., 2023) collection, which includes questions and relevant documents, and CHQ-Summ (Yadav et al., 2022), which contains consumer health questions. Using the retrieved documents from the MMR baseline approaches (both unimodal and multimodal), we developed two unimodal baselines, `Text-RRG` and `Video-RRG`. In these settings, the top-10 documents and videos from the MMR baselines were fed into the model to generate answers with citations using the prompts shown in Fig. 4. We also included a multimodal baseline, `MM-RRG`, that combines the top-5 textual and video sources obtained from the MMR baselines to generate answers with citations. We observed that video transcripts are often too long and may exceed the model's maximum token limit of 4096. Therefore, in `Video-RRG` and `MM-RRG`, we summarized the transcripts before feeding them to the model for answer generation. The prompt used for transcript summarization is shown in Fig. 5. We generated answers and summaries both using a sampling-based decoding strategy with a maximum output length of 350 tokens. Decoding was performed at a temperature of 0.7 and with nucleus sampling with $p = 0.9$.

4.2.3 VAL Baselines

We developed a baseline for the VAL task by adopting TimeLens (Zhang et al., 2025), a multimodal language model designed for video temporal grounding, built on top of Qwen2.5-VL-7B-Instruct model (Wang et al., 2024a). TimeLens introduces a simple interleaved textual prefix for time representation and is trained using a reinforcement learning with verifiable rewards (RLVR) approach. The training process employs practical strategies, including early stopping based on reward saturation

[9] https://github.com/castorini/pyserini

[10] cross-encoder/ms-marco-MiniLM-L-6-v2

Team Name	Team Affiliations	MMR	MAG	VAL
VAJRA-DOMINATORS	Pune Institute of Computer Technology, India	✓	✗	✗
PRIDE-BOILERS	Purdue University Northwest, USA	✓	✗	✗
SEAHAWK	Nankai University, China University of North Carolina Wilmington, USA	✗	✓	✓
LAMAR-2	Mahidol University, Thailand	✗	✗	✓
UNCC	University of North Carolina Charlotte, USA	✗	✗	✓
NJUST-KMG	Nanjing University of Science and Technology, China	✓	✓	✓
ADAPT	NA	✗	✗	✓
405621	NA	✗	✗	✓

Table 1: Participating teams and their task participation at MedGenVidQA 2026 shared task. The teams that did not disclose their affiliations are shown as NA.

and difficulty-aware data sampling. The model was trained on the TimeLens-100K dataset, specifically designed for temporal grounding tasks. We utilized the TimeLens-7B model with its default generation and video processing settings to generate the start and end timestamps of the answer, using the prompt shown in Fig. 6.

We provided the processed datasets and baseline implementations as a starter kit[11] for participants to build upon their approaches.

5 Participating Teams and Methods

5.1 Participating Teams

We used the Codabench platform to release the datasets, and for registration, and submissions of the participating teams. In total, 8 teams from Asia (India, China, Thailand), and North America (USA) participated in the MedGenVidQA 2026 shared task and submitted a total of 46 runs. We have provided (*cf* Table 1) the team name, affiliations, and their participation in MRR, MAG, and VAL tasks. For Task MRR and MAG, we did not perform real-time evaluation on Codabench because the tasks require expert evaluation of the runs. The results of all the participating teams for the VAL task[12] are available on the Codabench platform.

5.2 MMR Submissions

5.2.1 Methods

VAJRA-DOMINATORS (Dhaktode et al., 2026) proposed a multi-stage hybrid retrieval approach over both the textual and video corpus. For document retrieval, queries were processed to extract medical keywords that were submitted to the NCBI PubMed E-utilities (ESearch) API[13] using a sequence of fallback strategies, including full-query, medical keyword-based, and truncated queries, with results ranked using a normalized scoring scheme based on an exponential decay function: $Score = 0.5 + 0.49 \times e^{(-rank \times 0.3)}$. For video retrieval, the system combined dense retrieval using a PubMedBERT bi-encoder (trained with hard negatives) and sparse BM25 search over temporally segmented transcripts. The outputs from both stages were fused and further refined using a biomedical cross-encoder for re-ranking, with deduplication applied at the video level. A T5-based model was used for query expansion to generate clinical synonyms and improve matching between user queries and technical content. The overall approach focused on improving recall while maintaining relevance through re-ranking and query refinement.

PRIDE-BOILERS (Ebinesar et al., 2026) proposed CRAG-MMR, a five-stage multimodal retrieval pipeline that addresses the vocabulary mismatch between lay health queries and clinical literature. For document retrieval, a BM25 index was built over the PubMed collection, while dense retrieval was performed using MedCPT (ncbi/MedCPT-Article-Encoder) encoders, with documents represented as 768-dimensional embeddings stored in a FAISS index (Douze et al., 2025). To better capture different aspects of the information needed, each query was decomposed into multiple clinically oriented sub-queries using Gemini Pro, targeting treatment options, procedures, prevention strategies, and guidelines. In addition, a Hypothetical Document Embedding (HyDE) (Gao et al., 2023) based approach was employed, in which the Gemini Pro generated a synthetic abstract that was encoded and used as an additional dense query. The resulting BM25, dense, and HyDE retrieval outputs were combined using the Reciprocal Rank Fusion (RRF) strategy. For video retrieval, subtitles from instructional videos were indexed using both BM25 and SapBERT-based dense representations. Queries were adapted to better match the spoken and procedural nature of video content, and retrieval was performed independently over both signals. The results were fused using the RRF strategy for document retrieval.

[11] https://github.com/medgenvidqa/starter-kit

[12] https://www.codabench.org/competitions/14015/

[13] https://www.ncbi.nlm.nih.gov/books/NBK25497/

Rank	Team	Run/Model	Textual						Videos					
			MAP	R@5	R@10	P@5	P@10	nDCG	MAP	R@5	R@10	P@5	P@10	nDCG
Baselines														
–	Baseline	`Video-RR`	–	–	–	–	–	–	**0.5884**	**0.6067**	**0.6528**	**0.41**	**0.2217**	**0.6616**
–	Baseline	`Text-RR`	0.5404	0.5505	0.5863	0.5133	0.27	0.646	–	–	–	–	–	–
Participants Systems														
1	PRIDE-BOILERS	CRAG-MMR	**0.5550**	**0.5571**	**0.5866**	**0.5333**	**0.2817**	**0.6532**	0.5304	0.5478	0.5833	0.3900	0.2100	0.5927
2	NJUST-KMG	Run1	0.0167	0.0167	0.0167	0.0200	0.0100	0.0194	0.0167	0.0167	0.0167	0.0067	0.0033	0.0167
3	VAJRA-DOMINATORS	Run1	0.0024	0.0024	0.0024	0.0033	0.0017	0.0048	0.0167	0.0167	0.0167	0.0033	0.0017	0.0167
		Run2	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0033	0.0017	0.0167
		Run3	–	–	–	–	–	–	0	0	0	0	0	0
		Run4	–	–	–	–	–	–	0.01	0.02	0.02	0.0067	0.0033	0.0141

Table 2: Performance of the participating teams on the **MMR** task considering the retrieved documents from the PubMed corpus and videos from the video corpus. Teams are ranked based on their best run using nDCG as the primary metric. "–" indicates that no results were returned for the corresponding modality (text or video) in the submitted run.

Systems	Correctness			Completeness		
	Auto (All)	Auto (10Q)	Nugget (10Q)	Auto (All)	Auto (10Q)	Nugget (10Q)
`Text-RRG`	0.6122	0.5356	0.8586	0.6578	0.6429	0.4338
`Video-RRG`	0.5982	0.5188	0.7764	0.5305	0.3482	0.3660
`MM-RRG`	0.6224	0.5461	0.7914	0.6324	0.3898	0.4014

Table 3: Performance comparison of baseline systems in terms of answer correctness and completeness on the MAG task. Automatic evaluation (Auto) using BioACE is reported over all 60 questions, while nugget-based (Nugget) evaluation was computed on a subset of 10 questions. For comparison between Nugget-based evaluation and Automatic evaluation, we provide the comparative results on the same set of 10 questions.

Systems	Citation Coverage	Citation Support Rate	Citation Contradict Rate
`Text-RRG`	**0.6667**	**0.8170**	0.0308
`Video-RRG`	0.2221	0.7482	**0.0073**
`MM-RRG`	0.3430	0.6863	0.0140

Table 4: Performance comparison of baseline systems in terms of citation quality using BioACE for the MAG task.

5.2.2 Results

We have provided the results for the MMR task in Table 2. The team PRIDE-BOILERS achieved the best results in retrieving textual documents from PubMed, with an nDCG of 0.6532. The team also achieved the highest nDCG score of 0.5927 among all participants' submitted runs; however, it was lower than the baseline `video-RR` approach. The baseline approaches show strong performance in both modalities, demonstrating the effectiveness of the BM25 followed by the re-ranking strategy over the LLM-assisted retrieval strategy.

5.3 MAG Submissions

Two teams, SEAHAWK and NJUST-KMG, submitted runs; however, they did not provide any details about them, and their submissions were incomplete, so we could not include them.

5.3.1 Results

We have evaluated the baseline runs using BioACE, and the results of the answer and citation evaluations are reported in Tables 3 and 4, respectively. For answer evaluation, BioACE provides automatic assessment based on the relevant literature, as well as nugget-based assessment, in which the ground-truth nuggets are used to compute the completeness and correctness of the generated answer. We constructed ground-truth nuggets for 10 questions in the test set and conducted both nugget-based and automatic evaluations on this subset, along with automatic evaluation over the entire test set. The results show that the multimodal baseline `MM-RRG` achieves the highest automatic correctness when evaluated over all questions, indicating the benefit of combining relevant textual and video sources. However, in nugget-based correctness evaluation, the `MM-RRG` model performs best; likely because the nuggets were constructed primarily from textual sources. `Text-RRG` obtained the highest scores across for answer completeness evaluation on both full and subset evaluations, while Video-RRG consistently underperforms, particularly on the 10-question subset. Overall, the results highlight that while multimodal approaches improve general answer quality, textual retrieval remains more reliable for fine-grained factual completeness

Rank	Team	Run/Model	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
Baseline						
–	Baseline	TimeLens-7B	78.75	63.75	48.75	61.09
Participants Systems						
1	LAMAR-2	Best Run	93.75	90.00	**77.50**	79.55
2	NJUST-KMG	Best Run	92.50	81.25	67.50	75.48
3	405621	Best Run	60.00	55.00	47.50	50.78
4	SEAHAWK	Best Run	71.25	52.50	42.50	52.30
5	UNCC	Best Run	62.50	36.25	22.50	42.57
6	ADAPT	Best Run	10.00	10.00	8.75	8.62

Table 5: Performance of the participating teams on the **VAL** task. Teams are ranked based on their best run using IoU@0.7 as the primary metric.

and correctness. To examine the effectiveness of the BioACE automatic evaluation framework, we computed rank correlations between Auto (10Q) and Nugget (10Q) scores using Spearman's ρ rank correlation coefficient. For correctness, the results show moderate agreement ($\rho = 0.50$), indicating some variation in system ranking. In contrast, completeness shows perfect agreement ($\rho = 1.0$), suggesting strong alignment between automatic and nugget-based evaluation for answer completeness.

For citation evaluation, we used the BioACE framework. The results show that `Text-RRG` achieves the highest citation coverage and support rate, while `Video-RRG` yields the lowest contradiction rate. We observed a notable drop in coverage for both `Video-RRG` and `MM-RRG`. This may be due to the fact that the answer generation model was fine-tuned to generate citations from PubMed-style textual documents rather than from video transcripts.

5.4 VAL Submissions

5.4.1 Methods

LAMAR-2 (Sermsrisuwan et al., 2026) utilizes a multimodal, multi-stage pipeline designed for precise temporal localization in medical videos. The proposed approach begins with parallel audio and visual extraction. For the audio stream, they extract 16kHz mono audio using FFmpeg and process it through `Qwen3-ASR-1.7B` with a forced aligner to generate exact word-level transcriptions. For the visual stream, they used PySceneDetect[14] to detect natural camera cuts and extract individual video segments, which were then passed to `Qwen3-VL` to generate chronological textual descriptions of the physical actions. Both the word-level transcripts and the scene-by-scene visual context were then temporally aligned and aggregated into a unified reference dataset for each video. In the final stage, they fed the raw video file, the user question, and this reference transcript and context into `Gemini-3-Flash`. They employed a vision-anchored prompting strategy, instructing the model to treat the aggregated text purely as background reference. Rather than relying on textual alignment, this instruction explicitly forces the model to ground its temporal predictions in the physical movements and procedures directly observed in the video footage, ultimately outputting the final starting and ending markers in a structured JSON format.

The team NJUST-KMG's (Li and Yang, 2026) approach to the VAL task began with ASR-based text extraction, followed by LLM-driven generation of multi-scale (single- and multi-sentence) semantic priors, which were fused with dense visual features. They used dynamic span queries for boundary regression and optimized the network using the 1D-GIoU (Rezatofighi et al., 2019) loss.

SEAHAWK (Tian and Dogan, 2026) utilized the timestamped video transcripts obtained from the Whisper (Radford et al., 2023) model to use in their answer localization system. They developed a retrieval-and-selection pipeline that treats visual answer localization as segment-level paragraph selection from timestamped video transcripts. Given a question and a segmented transcript, the system prompts DeepSeek (Liu et al., 2024) to select a contiguous range of transcript segments rather than directly generating timestamps.

UNCC (Demirhan and Zadrozny, 2026) employed a multi-stage pipeline for the VAL task. For each video-question pair, the timestamped transcript segments were first extracted using `GPT-4o Transcribe Diarize`, forming a structured

[14] https://www.scenedetect.com/

representation of the video content. Based on the query, candidate answer intervals were proposed from these segments. These candidates were then evaluated and refined using `GPT-5.4`, which considered both the transcript context and sampled key video frames. The pipeline further applied contextual augmentation and consistency-aware span selection, followed by a final refinement step to determine the start and end timestamps for the output.

5.4.2 Results

The VAL results in Table 5 show that the team LAMAR-2 achieved the highest IoU performance across multiple thresholds and the mIOU metric. It outperformed the baseline systems in terms of IoU@0.7, increasing from 48.75 to 77.50. Their strong performance is due to fine-grained modeling of the modalities (audio, visual), explicitly leveraging visual cues rather than relying on textual descriptions of the videos. The team NJUST-KMG achieved the second-best performance by using LLM-driven generation of multi-scale (single- and multi-sentence) semantic priors. The team UNCC achieved a competitive IoU@0.3. However, their performance declined for strict metrics such as IoU@0.5 and IoU@0.7. This may be because their approaches rely on textual candidate generation and on varying visual cues and temporal alignment. The baseline system achieved competitive performance, with an IoU@0.7 score of 48.75, outperforming the best approaches from four participants. We find that approaches that use multimodal grounding and temporal alignment achieve better results than those that rely solely on textual descriptions.

6 Discussion

For the MMR task, we found that the BM25-based approach is a strong method for extracting an initial set of relevant videos from the PubMed Central video transcripts. Hybrid approaches that combine lexical and dense retrieval show improved performance. We observed that video collections sourced from YouTube contain very few relevant videos for professional queries, particularly those involving medical procedures. To address this limitation, we plan to enrich the collection with more specialized videos and, in future work, explore open retrieval from a broader range of video sources. We were also unable to evaluate generative retrieval (Li et al., 2025) and the joint retrieval setup (video and document projection in the unified space) in the current challenge, and we leave these to future work.

We note that the MAG task received limited participation. This is attributed to the complexity of the MAG task, which requires retrieving and synthesizing textual and video evidence to generate coherent and informative answers to medical questions. However, we provide baseline approaches and evaluation metrics to support further research on this important and timely task for the research community. It is to be noted that for the MMR and MAG baseline approaches, we use video information in the form of transcripts, which do not capture temporal dynamics. Despite this limitation, the models still achieve competitive performance and serve as reference points to guide further research on incorporating and enhancing visual information in retrieval and answer generation approaches.

The VAL task results show that the majority of approaches struggle to identify precise video segments that can be considered as answers. However, methods that incorporate fine-grained integration of multimodal information and alignment tend to achieve better performance, as observed in some of the submitted runs. Further analysis of the VAL ground-truth dataset reveals that many annotated segments are relatively long and correspond to many procedural or instructional steps. The developed test questions tend to be more general than focusing on a specific step. In future work, we plan to design more fine-grained video-centric questions and annotate shorter, more precise segments to better evaluate the models' ability to localize exact answer spans.

The results of this shared task highlight several practical challenges in developing medical multimodal question answering. The textual sources play a critical role in providing reliable answers. The video information is difficult to use effectively, even though it provides key information for medical professionals, including procedures and medical education. Specifically for the answer generation task, synthesizing answers from multimodal sources and ensuring that they remain attributable to the corpora is challenging. Moreover, evaluating multimodal medical question answering systems remains challenging because it relies on costly and time-intensive expert annotation. BioACE offers a useful alternative for efficient evaluation and ranking of system outputs and aligns well with nugget-based evaluation for completeness, but less so for correctness.

7 Conclusion

This paper provides an overview of MedGenVidQA 2026 shared task, organized as part of the BioNLP 2026 workshop. We discussed the tasks, datasets, evaluation metrics, and baseline systems. We also provided a summary of the approaches developed by the participating systems. We observed that efficient use of multimodal information sources remains challenging for the related tasks. Although multimodal approaches show promise, they do not consistently outperform strong textual baselines, suggesting that current fusion strategies remain limited. We are optimistic that introducing these tasks, baseline systems, and datasets will foster research toward reliable and trustworthy medical question answering systems for consumers and medical professionals.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Souvik Chowdhury and Badal Soni. 2025. R-vqa: A robust visual question answering model. *Knowledge-Based Systems*, 309:112827.

Hilmi Demirhan and Wlodek Zadrozny. 2026. UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Multimodal Medical Video Question Answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Pratik Vijay Dhaktode, Suhani Bighane, and Anupama Phakatkar. 2026. Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Basil Ebinesar, Keyuan Jiang, Charansai Maddineni, and Ashok Raja. 2026. Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2024a. Towards answering health-related questions from medical videos: Datasets and approaches. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025a. a dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2026. Bioace: An automated framework for biomedical answer and citation evaluations.

Deepak Gupta and Dina Demner-Fushman. 2022. Overview of the medvidqa 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274.

Deepak Gupta and Dina Demner-Fushman. 2024. Overview of trec 2024 medical video question answering (medvidqa) track.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024b. Overview of TREC 2024 Biomedical Generative Retrieval (BioGen) Track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*, NIST Special Publication. National Institute of Standards and Technology (NIST).

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2025b. Overview of TREC 2025 Biomedical Generative Retrieval (BioGen) Track. In *The Thirty-Fourth Text REtrieval*

Conference Proceedings (TREC 2025), NIST Special Publication. National Institute of Standards and Technology (NIST).

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.

Jinglong Li and Yang Yang. 2026. NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 109–119.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. Multitabqa: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 6322–6334.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

Watcharitpol Sermsrisuwan, Nopporn Lekuthai, Seksan Yoadsanit, and Titipat Achakulvisut. 2026. LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Xiaotian Tian and Gulustan Dogan. 2026. Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization. In *Proceedings of the 25th Workshop on Biomedical Language Processing (Shared Tasks)*, San Diego, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and 1 others. 2024b. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. 2025. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*.

Liang Zhang, Anwen Hu, Jing Zhang, Shuo Hu, and Qin Jin. 2023. Mpmqa: multimodal question answering on product manuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13958–13966.

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.
Documents:
{Document 1}
{Document 2 }
...
{Document 10}
Question : {question}

Figure 4: Prompt used for answer generation with citations.

Instruction: Write an accurate and concise summary of the provided document in relation to the given question. Focus on the information that directly helps answer the query and ignore irrelevant details. Use an unbiased and journalistic tone. Base the summary strictly on the document and do not add external knowledge.
Document: {document}
Question : {question}

Figure 5: Prompt used to summarize the video transcript.

Instruction: You are given a full video and the question: '{query}'. Locate the single continuous video segment in the video that best answers the question. Return exactly one start time and one end time for the most relevant segment. Choose the shortest segment that contains enough evidence to answer the question. If multiple candidate segments exist, return only the best one. If the video does not contain an answer, return 'none'. Output exactly in this format: 'The answer segment is <start time> - <end time> seconds' or 'none'.

Figure 6: Prompt used to locate the visual answer from the medical professional video.

Overview of the ClinicalSkillQA 2026 Shared Task on Continuous Perception and Procedural Reasoning in Clinical Skill Assessment

Xiyang Huang[1,2*] **Renxiong Wei**[3*] **Yihuai Xu**[1,2] **Zhiyuan Chen**[1,2] **Keying Wu**[1,2]
Jiayi Xiang[1,2] **Buzhou Tang**[4] **Yanqing Ye**[3] **Jinyu Chen**[3] **Cheng Zeng**[1]
Min Peng[1,2,†] **Qianqian Xie**[1,2,†] **Sophia Ananiadou**[5]

[1]School of Artificial Intelligence, Wuhan University
[2]Center for Language and Information Research, Wuhan University
[3]Zhongnan Hospital of Wuhan University
[4]Harbin Institute of Technology, Shenzhen [5]The University of Manchester

Abstract

This paper presents an overview of the ClinicalSkillQA 2026 shared task, which was organized with the BioNLP Workshop at ACL 2026. The goal of this shared task is to evaluate continuous perception and procedural reasoning in clinical skill assessment by requiring systems to reconstruct the correct temporal order of shuffled clinical key frames and generate rationales grounded in clinical workflow knowledge. The benchmark contains 200 test-only instances sampled from clinical skill videos, covering three emergency-care procedures. Each instance is annotated with the ground-truth temporal order and an expert-verified rationale. A total of seven teams participated in the task, collectively making 90 submissions, with four teams providing system description papers. Systems are evaluated using Task Accuracy, Pairwise Accuracy, and BERTScore, which measure exact sequence reconstruction, local temporal consistency, and rationale quality, respectively. In this paper, we describe the task setup, dataset construction, and evaluation criteria. We further summarize the methodologies adopted by participating teams and present a comprehensive analysis of the submitted systems. The official results suggest that current models still struggle with continuous perception and procedural reasoning, especially when they must integrate visual evidence, temporal structure, and clinical workflow knowledge.

1 Introduction

Clinical skill assessment is a core component of medical education, where trainees are evaluated on whether their procedures follow standardized clinical workflows and safety-critical operational requirements (Kogan et al., 2009, 2011). In scenarios such as cardiopulmonary resuscitation, automated external defibrillator operation, and bag-mask ventilation, expert assessors review trainees' operations and judge whether each action is performed in the correct procedural context (Harden et al., 1975; Regehr et al., 1998; Hodges et al., 1999). This process requires more than recognizing isolated visual events: assessors must understand how clinical actions unfold over time, how earlier steps constrain later ones, and how a sequence of observations corresponds to a valid medical procedure. As multimodal large language models (MLLMs) are increasingly explored for medical education and clinical decision support, it becomes important to evaluate whether they can connect visual evidence with procedural medical knowledge and provide explanations that are meaningful to clinical experts.

Recent benchmarks have advanced the evaluation of multimodal reasoning in both general and medical domains (Yue et al., 2024; Zuo et al., 2025; Yu et al., 2025; Xun et al., 2025; Wang et al., 2025b). However, many existing tasks focus on static image understanding (Liu et al., 2024b), multi-image question answering (Liu et al., 2024a), or general-domain temporal reasoning (Song et al., 2025), leaving clinical procedural understanding underexplored. Medical benchmarks such as MEDFRAMEQA (Yu et al., 2025) require reasoning over multiple medical images, but the questions can often be answered by aggregating evidence from individual frames, without explicitly modeling the temporal dependencies among clinical actions. Similarly, continuous perception benchmarks such as CP-Bench (Wang et al., 2025b) evaluate whether models can integrate informa-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1101–1108
July 3-4, 2026 ©2026 Association for Computational Linguistics

tion across visual observations, but often simplify continuous perception into tasks such as counting, where the temporal order of observations may not directly affect the final answer. These settings are insufficient for clinical skill assessment, where the interpretation of an action is inherently conditioned on its position within a procedural workflow.

This limitation becomes especially important in clinical training scenarios. For example, a frame showing chest compressions cannot be fully interpreted without considering whether patient assessment, emergency response activation, and preparation steps have occurred beforehand. Likewise, a frame showing defibrillator use only becomes clinically meaningful when placed within the correct sequence of rhythm assessment, pad placement, safety checks, and shock delivery. Thus, evaluating MLLMs for clinical skill assessment requires tasks that go beyond recognizing what appears in each frame and instead test whether models can reconstruct temporally coherent and clinically plausible action progressions. It also requires explanation-based evaluation, since in medical education a model's answer is useful only when its reasoning can be inspected and aligned with clinical workflow knowledge.

To address this challenge, we organize the ClincalSkillQA shared task for evaluating continuous perception and procedural reasoning in clinical skill assessment. The task asks MLLMs to arrange shuffled key frames into the correct temporal order of clinical actions and to generate rationales explaining the predicted order. This formulation directly tests whether models can infer action progression from visual observations while grounding their decisions in clinical workflow knowledge. We provide 200 sets of shuffled key frames extracted from clinical skill videos covering three representative emergency-care procedures. Each sample is annotated with an expert-defined ground-truth ordering and a reference rationale. For evaluation, we use Task Accuracy to measure exact sequence reconstruction, Pairwise Accuracy to assess local temporal consistency between adjacent clinical actions (Wang et al., 2025a), and BERTScore (Zhang et al., 2019) to evaluate the semantic quality of generated rationales. ClinicalSkillQA provides a clinically grounded testbed for studying how MLLMs connect visual observations, temporal structure, and procedural medical knowledge in medical education scenarios.

2 Task Description

ClinicalSkillQA formulates clinical skill understanding and continuous perception in clinical skill assessment as a multi-image temporal ordering task. Given a set of shuffled key frames extracted from a clinical skill video, a model is required to arrange the frames into a coherent sequence of clinical actions and provide a textual explanation for the resulting order. The task is designed to evaluate whether MLLMs can infer temporal continuity, procedural dependencies, and action-state transitions from fragmentary visual observations, rather than treating each frame as an independent image.

2.1 Task Definition

For each instance, the input consists of a set of shuffled key frames:

$$\mathcal{I} = \{I_1, I_2, \ldots, I_n\},$$

where each frame is associated with a unique identifier, such as A, B, C, or D. The frames are sampled from a continuous clinical skill video and randomly permuted before being presented to the model. The model is asked to output the correct temporal order of the frame identifiers and to generate an explanation that justifies this order.

The ground-truth answer is an ordered sequence:

$$I^* = (I_1^*, I_2^*, \ldots, I_n^*),$$

where I_t^* denotes the frame identifier corresponding to the t-th step in the original clinical procedure. The model prediction is formulated as:

$$\hat{y} = f_\theta(\mathcal{I}, \mathcal{P}),$$

where f_θ denotes the evaluated model and $\mathcal{P}$ denotes the task instruction that specifies the required output format and asks the model to recover the temporal order of the shuffled frames with an explanation. The prediction $\hat{y}$ contains two components:

$$\hat{y} = (\hat{I}, \hat{R}),$$

where $\hat{I} = (\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_n)$ is the predicted temporal order of the provided frame identifiers, and $\hat{R}$ is the generated rationale explaining the temporal dependencies and procedural cues used to infer this order. A valid prediction should include each provided frame identifier exactly once. Predictions with missing, duplicated, or invalid identifiers are regarded as invalid outputs by the official evaluation script.

2.2 Evaluation Setting

The official shared task evaluates both the predicted ordering and the generated rationale. For the ordering results, we use Task Accuracy and Pairwise Accuracy. Task Accuracy measures whether the predicted sequence exactly matches the ground-truth order. Pairwise Accuracy measures the fraction of correctly ordered adjacent frame pairs, providing a more fine-grained assessment of local temporal consistency.

For rationale evaluation, we use BERTScore to measure the semantic similarity between the generated rationale and the expert-annotated reference rationale. Together, these metrics assess whether a model can recover the correct temporal order while also producing explanations that are consistent with clinical workflows and expert reasoning.

3 Dataset

ClinicalSkillQA is constructed from clinical skill assessment videos and consists of 200 test instances of shuffled key frames. Each instance corresponds to a continuous clinical action sequence and is associated with a hidden ground-truth temporal order and an order rationale. The dataset is designed as a test-only shared task benchmark to evaluate whether MLLMs can reconstruct coherent clinical procedures from fragmentary visual observations. As summarized in Table 1, ClinicalSkillQA covers three clinical skill scenarios and uses shuffled key frames as input, with temporal order and rationale as evaluation targets.

3.1 Data Source

The videos are derived from the SiMing-Bench (Huang et al., 2026) dataset, which contains medical student clinical procedure videos collected under standardized clinical skills assessment settings. We sample key frames from the videos to construct multi-image instances, where each instance represents a continuous clinical action sequence. This source provides realistic yet controlled procedural scenarios for evaluating continuous perception in medical contexts. Compared with independent medical image recognition, the sampled frames require models to infer temporal and procedural dependencies among clinically related actions, such as preparation, patient contact, device placement, action execution, and post-action checking.

The study was approved by the Institutional Review Board. During data collection and processing, privacy protection procedures were applied to remove or mask personally identifiable information. The dataset is intended for research on multimodal clinical skill understanding and medical education, rather than for direct clinical decision-making.

Item	Description
Dataset	ClinicalSkillQA
# Instances	200
Scenarios	CPR, AED, BMV
Input	Shuffled key frames
# Frames	4 or 6 per instance
Target	Order & rationale
Evaluation setting	Test-only benchmark

Table 1: Dataset summary of ClinicalSkillQA. Each instance consists of 4 or 6 shuffled key frames sampled from a continuous clinical skill video. The hidden temporal order and expert-verified rationale are used for official evaluation.

3.2 Annotation

Each instance is annotated with two types of supervision: the ground-truth temporal order and an order rationale. The temporal order is derived from the original chronological sequence of the sampled key frames. The rationale explains why the recovered order is clinically and procedurally correct.

We adopt a model-assisted and physician-verified annotation pipeline to improve both efficiency and consistency. Specifically, a multimodal model[1] is first prompted with the shuffled key frames and the corresponding ground-truth order to generate an initial order rationale. The prompt asks the model to focus on visible action progression, state transitions, and clinically meaningful dependencies across frames. The generated rationale is then refined by a second multimodal model[2] to improve clarity, remove redundant descriptions, and ensure that the explanation is aligned with the visual evidence.

Finally, two physicians independently review the temporal order and the refined rationale. They correct clinically inaccurate statements, remove unsupported visual claims, and verify that the rationale reflects the actual procedural progression shown in the frames. Disagreements are resolved through discussion until consensus is reached. The final annotations are used only for official evaluation and

[1]We used Gemini 3 Pro for the first-stage rationale generation.

[2]We used GPT-5.2 for the second-stage rationale refinement.

are not released to participants during the shared task.

4 Evaluation

ClinicalSkillQA evaluates model performance from two perspectives: the correctness of the predicted temporal ordering and the quality of the generated rationale. Following the task definition, each prediction contains an ordering result $\hat{I}$ and an ordering rationale $\hat{R}$. We use Task Accuracy and Pairwise Accuracy to evaluate the predicted ordering, and BERTScore to evaluate the rationale.

4.1 Task Accuracy

Task Accuracy measures whether the model reconstructs the entire temporal sequence exactly. For the i-th instance, let the ground-truth ordering be

$$I^{*(i)} = (I_1^{*(i)}, I_2^{*(i)}, \ldots, I_{n_i}^{*(i)}),$$

and the predicted ordering be $\hat{I}^{(i)}$. The prediction is counted as correct only when the two sequences are identical:

$$\mathrm{Acc}_i = \mathbb{I}\left(\hat{I}^{(i)} = I^{*(i)}\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Given N test instances, Task Accuracy is computed as

$$\mathrm{TaskAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{I}^{(i)} = I^{*(i)}\right).$$

This metric is strict: any misplaced, missing, or duplicated frame identifier causes the instance to be marked as incorrect.

4.2 Pairwise Accuracy

Task Accuracy does not reflect partial temporal correctness. Therefore, we further report Pairwise Accuracy, which measures whether the model preserves the correct relative order of adjacent frame pairs in the ground-truth sequence.

For the i-th instance, let $\mathrm{pos}_{\hat{I}^{(i)}}(x)$ denote the position of frame identifier x in the predicted ordering. For each adjacent ground-truth pair $(I_t^{*(i)}, I_{t+1}^{*(i)})$, we define

$$c_{i,t} = \mathbb{I}\left(\mathrm{pos}_{\hat{I}^{(i)}}(I_t^{*(i)}) < \mathrm{pos}_{\hat{I}^{(i)}}(I_{t+1}^{*(i)})\right).$$

If a required frame identifier is missing, the corresponding pair is treated as incorrect.

We compute Pairwise Accuracy by aggregating all adjacent ground-truth pairs across the evaluation set:

$$\mathrm{PairAcc} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i-1} c_{i,t}}{\sum_{i=1}^{N}(n_i - 1)}.$$

This metric only checks whether each adjacent ground-truth pair appears in the correct relative order in the prediction; the two frames do not need to remain adjacent in the predicted sequence. Pairwise Accuracy provides partial credit for recovering local temporal transitions, even when the full sequence is not perfectly reconstructed.

4.3 BERTScore for Rationale Evaluation

We also evaluate the generated ordering rationale $\hat{R}$ against the expert-verified reference rationale R^*. BERTScore is used because it measures semantic similarity based on contextual embeddings rather than exact lexical overlap. This is suitable for ClinicalSkillQA, where different explanations may use different wording but still describe the same clinical operation, procedural stage, visual evidence, and temporal transition. We report the average BERTScore F1 over all test instances.

4.4 Overall Score

The official leaderboard score aggregates temporal ordering performance and rationale quality. We first define the ordering score as

$$S_{\mathrm{ord}} = \alpha \cdot \mathrm{TaskAcc} + (1-\alpha)\cdot \mathrm{PairAcc},$$

where α controls the relative importance of exact sequence reconstruction and partial temporal consistency.

The final score is then computed as

$$S_{\mathrm{final}} = 100 \cdot [\beta \cdot S_{\mathrm{ord}} + (1-\beta)\cdot \mathrm{BERTScore}_{F1}],$$

where β balances the ordering result and the generated rationale. In our official evaluation, we set $\alpha = 0.7$ and $\beta = 0.8$. This choice assigns higher importance to exact sequence reconstruction while still rewarding models for preserving local temporal transitions and generating semantically meaningful rationales. The factor 100 scales the score to a more readable range.

5 Results

We used the CodaBench platform[3] to manage the submission and evaluation process for the shared

[3] https://www.codabench.org/competitions/14884/#/pages-tab

task. Overall, seven teams participated and made a total of 90 submissions.

Table 2 reports the official results[4] on the ClinicalSkillQA test set. Teams are ranked by the overall score, which jointly considers exact temporal ordering, local temporal consistency, and rationale quality.

Overall, ZZUNLP achieves the best performance, with an overall score of 71.43. It also obtains the highest Task Accuracy and Pairwise Accuracy, reaching 0.63 and 0.86, respectively, and achieves a strong BERTScore F1 of 0.79. The second-ranked system, qqpprun, obtains an overall score of 63.88, with Task Accuracy and Pairwise Accuracy close to those of ZZUNLP. This indicates that the leading systems are relatively effective at reconstructing temporal sequences and preserving local temporal transitions.

The middle-ranked systems, including FBK and nance, achieve overall scores of 59.45 and 56.62, respectively. Although their ordering metrics are lower than those of the top two systems, they obtain strong BERTScore F1 scores. In particular, nance reaches a BERTScore F1 of 0.79, matching that of ZZUNLP. This suggests that rationale generation can remain competitive even when exact temporal reconstruction is less accurate.

The remaining systems obtain substantially lower overall scores, mainly due to limited Task Accuracy. For example, VerbaNexAI, DLNLP, and zhekun achieve Task Accuracy scores of 0.17, 0.10, and 0.03, respectively, while their Pairwise Accuracy scores remain higher. This gap indicates that these systems can recover partial local temporal relations but often fail to reconstruct the complete sequence exactly. Notably, zhekun obtains a relatively high BERTScore F1 of 0.74 despite its low ordering accuracy, further showing that rationale similarity does not necessarily imply correct temporal ordering.

Across all teams, Pairwise Accuracy is consistently higher than Task Accuracy. This trend shows that preserving local temporal relations is generally easier than reconstructing the full shuffled sequence exactly. Meanwhile, BERTScore F1 is relatively high for several systems even when their Task Accuracy remains limited. This suggests that rationale quality and ordering correctness capture complementary aspects of model performance: a system may generate clinically plausible explanations without fully recovering the correct temporal order. Overall, these results demonstrate that ClinicalSkillQA remains challenging for current systems, especially in requiring models to jointly perform fine-grained temporal reconstruction and clinically grounded rationale generation.

[4]Component metrics are rounded to two decimal places, while overall scores are computed using unrounded official metric values.

6 System Descriptions

This section summarizes the submitted systems for the ClinicalSkillQA shared task. Out of the seven participating teams, four teams submitted system papers. Overall, participating teams explored different strategies for reconstructing the temporal order of clinical keyframes and generating corresponding rationales. The submitted methods mainly relied on multimodal large language models, with different designs for visual evidence extraction, procedural reasoning, and explanation generation.

ZZUNLP. ZZUNLP proposed Perceive-and-Plan, a two-stage in-context learning framework for clinical skill keyframe reordering. The system first performs structured visual perception by generating compact clinical descriptions for each frame, enhanced with saliency-guided picture-in-picture inputs that zoom in on critical head/airway and chest/hand regions. It then conducts temporal planning in a fresh conversation, using BLS procedural rules and visual anchors to verify the predicted chronological order. Without parameter updates, the system achieved an overall score of 71.43, with 0.63 Task Accuracy, 0.86 Pairwise Accuracy, and 0.79 BERTScore F1. The results suggest that separating visual perception from temporal reasoning improves local ordering reliability, while the remaining gap between exact sequence accuracy and pairwise accuracy indicates that long-range permutation errors and subtle state-transition ambiguities remain challenging.

FBK. FBK adopted a training-free prompting framework based on Qwen3-VL-32B-Instruct for clinical keyframe ordering. Their system examined several prompt formulations, random filename renaming to reduce alphabetical-order bias, few-shot demonstrations, and keypoint-derived predicates extracted by YOLOv8 pose estimation. The strongest setting used randomly renamed frames and full human-pose predicates without few-shot examples, achieving an overall score of 59.45, with 0.47 Task Accuracy, 0.72 Pairwise Accuracy, and

Team	Overall Score	Task Accuracy	Pairwise Accuracy	BERTScore F1
ZZUNLP	71.43	0.63	0.86	0.79
qqpprun	63.88	0.58	0.84	0.55
FBK	59.45	0.47	0.72	0.76
nance	56.62	0.41	0.75	0.79
VerbaNexAI	37.96	0.17	0.60	0.71
DLNLP	30.44	0.10	0.55	0.57
zhekun	28.71	0.03	0.51	0.74

Table 2: Official results on the ClinicalSkillQA test set. Teams are ranked by the overall score. Pairwise Accuracy denotes micro-averaged pairwise accuracy.

0.76 BERTScore F1. Their analysis shows that simple prompting was more effective than more complex intermediate reasoning prompts, while filename renaming consistently improved performance by reducing spurious textual biases. The largest gain came from keypoint-derived predicates, suggesting that explicit posture and body-motion cues provide useful temporal evidence for reconstructing clinical action sequences.

VerbaNexAI Lab. VerbaNexAI Lab proposed a zero-shot two-stage framework based on Qwen2-VL-2B-Instruct. The system first extracts frame-level visual evidence and maps it into structured procedural states using deterministic rules. It then assigns each frame an ordinal stage score and sorts frames accordingly, with PCA over multimodal embeddings used to break ties. The final prediction includes both the ordered frame sequence and an evidence-based rationale. On the official test set, the system achieved an overall score of 37.96, with 0.17 Task Accuracy, 0.60 Pairwise Accuracy, and 0.71 BERTScore F1. These results indicate that the method produced complete and semantically reasonable rationales, but still struggled with exact temporal ordering, especially when adjacent procedural states were visually similar.

DLNLP. DLNLP proposed EvidenceFlow, a structured zero-shot framework built on Qwen2.5-VL-7B-Instruct for clinical keyframe ordering. The method decomposes the task into global overview, local evidence modeling, and ordering decision, and explores two variants: EvidenceFlow-M, which relies on model-led reasoning, and EvidenceFlow-R, which introduces rule-guided correction with explicit evidence constraints. On the official test set, the submitted EvidenceFlow-R variant achieved an overall score of 30.44, with 0.10 Task Accuracy, 0.55 Pairwise Accuracy, and 0.57 BERTScore F1. In their system paper, DLNLP further reports that EvidenceFlow-M achieved a higher BERTScore F1 of 0.74, whereas EvidenceFlow-R performed better on ordering-related metrics. These results suggest that explicit rule constraints improve ordering stability, whereas more flexible model-led reasoning may produce semantically richer explanations. The main remaining bottleneck is the stable integration of subtle cross-image temporal cues into a globally consistent clinical sequence.

7 Conclusion

We introduced ClinicalSkillQA, a BioNLP shared task for evaluating continuous perception and procedural reasoning in clinical skill assessment. Given shuffled key frames from clinical skill videos, systems are required to recover the correct temporal order and generate rationales explaining their predictions. The benchmark contains 200 test-only instances covering three emergency-care procedures, with annotations consisting of the ground-truth temporal order and expert-verified rationales. Official results show that current multimodal systems can capture some local temporal relations and produce plausible explanations, but still face clear challenges in exact full-sequence reconstruction. These findings suggest that clinical procedural reasoning remains difficult for existing multimodal large language models, particularly when visual evidence, temporal dependencies, and clinical workflow knowledge must be jointly integrated. ClinicalSkillQA provides a clinically grounded testbed for future research on reliable multimodal reasoning in medical education.

8 Limitations

ClinicalSkillQA has several limitations. First, the benchmark currently includes 200 test-only instances from three emergency-care procedures, which may not fully represent the diversity of clinical skills, medical specialties, and assessment environments. Expanding the dataset to more procedures and clinical contexts would improve its

coverage and generalizability.

Second, the task is based on shuffled key frames rather than full videos. This formulation enables controlled evaluation of temporal ordering and procedural reasoning, but it does not fully capture motion dynamics, action duration, or fine-grained temporal transitions. Therefore, the benchmark primarily evaluates whether models can infer procedural order from representative visual observations, rather than their complete ability to understand continuous clinical videos.

Third, rationale quality is evaluated using BERTScore F1. Although this provides an automatic measure of semantic similarity, it may not fully reflect clinical correctness, visual grounding, or educational usefulness. Future work could introduce clinically informed rationale metrics or expert-based evaluation.

Finally, ClinicalSkillQA is designed as a test-only shared task benchmark. This setting supports fair evaluation of submitted systems, but it does not examine the effects of supervised training or domain adaptation. Future versions may include development data, additional procedures, and full-video assessment settings.

References

Ronald M Harden, Mary Stevenson, W Wilson Downie, and GM Wilson. 1975. Assessment of clinical competence using objective structured examination. *Br Med J*, 1(5955):447–451.

Brian Hodges, Glenn Regehr, Nancy McNaughton, Richard Tiberius, and Mark Hanson. 1999. Osce checklists do not capture increasing levels of expertise. *Academic Medicine*, 74(10):1129–1134.

Xiyang Huang, Jiawei Lin, Keying Wu, Jiaxin Huang, Kailai Yang, Renxiong Wei, Jiayi Xiang, Ziyan Kuang, Min Peng, Qianqian Xie, and 1 others. 2026. Siming-bench: Evaluating procedural correctness from continuous interactions in clinical skill videos. *arXiv preprint arXiv:2604.09037*.

Jennifer R Kogan, Lisa Conforti, Elizabeth Bernabeo, William Iobst, and Eric Holmboe. 2011. Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical education*, 45(10):1048–1060.

Jennifer R Kogan, Eric S Holmboe, and Karen E Hauer. 2009. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *Jama*, 302(12):1316–1326.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and 1 others. 2024a. Mibench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Glenn Regehr, Helen MacRae, Richard K Reznick, and David Szalay. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an osce-format examination. *Academic Medicine*, 73(9):993–997.

Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342.

Qineng Wang, Wenlong Huang, Yu Zhou, Hang Yin, Tianwei Bao, Jianwen Lyu, Weiyu Liu, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, and 1 others. 2025a. Enact: Evaluating embodied cognition with world modeling of egocentric interaction. *arXiv preprint arXiv:2511.20937*.

Zeyu Wang, Zhenzhen Weng, and Serena Yeung-Levy. 2025b. Continuous perception matters: Diagnosing temporal integration failures in multimodal models. *Preprint*, arXiv:2408.07867.

Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, and 1 others. 2025. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *arXiv preprint arXiv:2505.02064*.

Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. 2025. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and

Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

Author Index

Aal Abdulsalam, Abdulrahman, 531
Abbaszadeh, Asal, 326
Afnan, Afnan, 997
Agarwal, Rohit, 687
Agarwal, Shubham, 665
Aiersilan, Aizierjiang, 64
Aizawa, Akiko, 435
Al Zaabi, Adhari, 531
Alharbi, Ebrahim, 843
Amiri, Hadi, 760
Ananiadou, Sophia, 382, 932, 1101
Andrade, Daniel, 679
Anvekar, Tejas, 944
Aramaki, Eiji, 997
Arazkhani, Niloofar, 880
Arikawa, Andrea, 326
Asada, Masaki, 407
Au, Steven, 656
Awotoro, Ebenezer, 853

Bada, Michael, 519
Badash, Zvi, 292
Baddour, Moussa, 45
Barlow, Stephen, 20
Bartels, Davis, 812, 1080
Bechet, Frederic, 209
Bellucci, Cristiano, 595
Belmadani, Ikram, 209
Ben Atya, Hadas, 292
Bernaschi, Massimo, 595
Bethard, Steven, 31
Bhojani, Krushil, 104
Bittner, Madeline, 481, 718, 812
Bonesteve, Axel, 45
Bongo, Lars, 464
Borowiack, Ewa, 955
Buhnila, Ioana, 413, 687
Buñales, Hannah Grachiella, 1029

Cai, Zhenyang, 1056
Campbell, Collin, 1089
Celi, Leo, 760
Chang, Chia-Hsuan, 156
Chen, Jinyu, 1101
Chen, Ling, 932
Chen, Zhaoming, 932
Chen, Zhiyuan, 1101
Cheng, Chi-Tung, 235
Cheng, Gao, 1056
Cheng, Jiali, 760
Choi, Woojae, 695
Choudhury, Manan, 944
Chowdhury, Marzana, 605
Chowdhury, Suparno, 944
Chu, Bohao, 753, 911
Chua, Kervin Gabriel, 1029
Clausel, Marianne, 864
Cochran, Brent, 880
Colelough, Brandon, 812, 1080
Comeau, Donald, 475, 502
Constant, Mathieu, 413, 687, 864
Coubez, Xavier, 864
Coulet, Adrien, 791
Cui, Haochen, 426
Cytter-Kuint, Ruth, 292

Dahal, Rajashree, 490
Dameron, Olivier, 45
Dao, An, 435
Davies, Joe, 20
De Dieuleveult, Maud, 791
De Tayrac, Marie, 45
Demberg, Vera, 426
Demner-Fushman, Dina, 481, 718, 812, 1080, 1089
Deshwal, Sarthak, 282
Diller, Matthew, 718
Djuric, Milosh, 342
Dobson, Richard, 191, 665, 992
Dong, Bin, 392
Dube, Nitish, 282
Dufour, Richard, 209

Egami, Shusaku, 407
El Keraby, Habiba, 531
El Khettari, Oumaima, 209
Elgaar, Mohamed, 760
Elmofty, Mohamed, 508
Eyre, David, 115, 426

Favre, Benoit, 209
Floresca, Rez Samantha, 1029
Focht, Gili, 292
Freiman, Moti, 292
Fuhr, Norbert, 753, 911
Fukuda, Ken, 407

Gagnier, Henry, 552
Gan, Yujian, 20
Gao, Rena, 932
Gavrielov, Naama, 292
Giuffrè, Mauro, 156
Golnari, Pedram, 1089
Gomez Mesa, Lina, 617
Grosdidier, Solène, 791
Guo, Xiaorui, 382
Guo, Yike, 605
Gupta, Ashok, 605
Gupta, Deepa, 605
Gupta, Deepak, 1089
Gupta, Vibhor, 605
Gupta, Vivek, 944

Hagopian, Talar, 292
Halabi, Mouhammad, 342
Han, Wooseok, 572
Han, Yikun, 630
Hanken, Keno, 742
Hao, Edric Castel, 1029
Hassan, Moustafa, 1
Hattab, Georges, 853
Henry, Sam, 199
Holgate, Ben, 20
Hong, Lian, 679
Hosseinpour, Pardis, 490
Hua, Yining, 932
Huang, Xiyang, 1101
Hunter, Lawrence, 519

Iacomini, Chiara, 595
Ilgen, Bahar, 853
Isayev, Olexandr, 268
Ishigaki, Tatsuya, 407
Islamaj, Rezarta, 801

Jadhav, Kshitij, 981
Jalan, Pratik, 981
Jeeballah, Riham, 531
Ji, Shaoxiong, 932
Ji, Yuting, 605
Ji, Ziqing, 729
Jiang, Shanshan, 392
Jiang, Yuhang, 920
Jo, Eunjung, 695
Jobanputra, Mayank, 426
Johnson, Alistair, 115
Jones, Athlene, 644
Joshi, Mukul, 981
Jung, Jisoo, 572
Just, Sandra, 464

Kabbani, Loay, 342
Kahanda, Indika, 326, 644
Kam, Tae-Eui, 695
Kamishetty, Pranithi, 490
Kang, Bogyeong, 695
Kao, Hung-Yu, 235
Kavuluru, Ramakanth, 920
Kenmotsu, Toshimune, 679
Khakwani, Kaneez, 944
Khamfroush, Hana, 970
Khan, Muhammed, 944
Khatun, Ayesha, 31
Kilicoglu, Halil, 630, 770
Kim, Gibaeg, 572
Kim, Minkyu, 572
Kim, Won, 475
Kim, Won Gyu, 502
Kirubakaran, Ashwin, 552
Kona, Chandralekha, 426
Kondadadi, Rishik, 135
Kotecki, Maciej, 880
Kumar, Gaurav, 704
Kwegyir-Afful, Ernest, 128

Labbe, Thomas, 45
Lan, Mengfei, 630
Lanz, Vojtech, 559
Laparra, Egoitz, 31
Le Thanh, Duy, 75
Leaman, Robert, 801
Lee, Deok-Joong, 695
Lee, Hyunkyung, 572
Lee, Jeonguk, 572
Lee, Sangwoon, 695
Leser, Ulf, 75, 508
Li, Ping-Chien, 235
Liao, Chien-Hung, 235
Liao, Zicen, 250
Lieu, Khanh, 644
Lilli, Livia, 595
Lin, Sylvey, 770
Lin, Thomas, 115
Lin, Ying-Jia, 235
Liu, Jinghui, 353, 372
Liu, Weichen, 268
Liu, Zhiwei, 382
Lo, Tzu-Chin, 235

Loftis, Richard, 326
Lu, Zhiyong, 475, 502, 801
Ly, Nhan, 435
Lynam, Jack, 199

Ma, Yixian, 753
Madeddu, Francesco, 595
Magotra, Akhilesh, 981
Mahanipour, Afsaneh, 970
Majkic, Ninoslav, 665
Manrique, Ruben, 617
Martinez Novoa, Santiago, 617
Martínez, Raquel, 84
Masciocchi, Carlotta, 595
Matsumoto, Yuji, 435
Maury, Fabien, 791
Menke, Joseph, 770
Ming, Shufan, 770
Miskov-Zivanov, Natasa, 268, 880
Mizogaki, Masato, 679
Mohtarami, Mitra, 760
Mok, Monica, 605
Moller-Grell, Niko, 665, 992
Montalvo, Soto, 84
Morales-Sánchez, Rodrigo, 84

Na, Hongbin, 932
Nam, Dongin, 770
Nam, Hyeonyeong, 695
Nandyala, Hima Bindu, 104
Naseer, Hulma, 426
Nguyen, Anthony, 353, 372
Nguyen, Michelle, 1014
Nigam, Shubham, 1041
Nishiyama, Tomohiro, 997
Nourbakhsh, Erfan, 890
Nowak, Artur, 955

Ondov, Brian, 156
Opa?ek, Monika, 955
Ortega, John E., 135
Oza, Samarth, 426
Ozler, Kadir Bulut, 31

Pal, Debmalya, 1066
Pamulaparthy, Satwik, 490
Parde, Natalie, 490
Patarnello, Stefano, 595
Patel, Piyush, 1041
Pecina, Pavel, 559
Pelloni, Olga, 464
Peng, Min, 1101
Prasad, Dilip K., 687
Prieto, Juan, 617
Przechera, Bartosz, 955
Purver, Matthew, 250

Qu, Xiaodong, 64
Quispe, Rodolfo, 115

Ramisch, Carlos, 209
Ratas, Mart, 191
Remy, François, 713
Reyes, Georgianna, 1029
Riaz, Irbaz, 944
Richardson, Mark, 20
Rios, Anthony, 890
Roberts, Angus, 181
Rogers, Willie, 718
Rollier, Paul, 45
Roy, Mahule, 15
Roy, Subhas, 15
Rubinfeld, Ilan, 342

Sadowska, Ewelina, 955
Sadrolashrafi, Mehrnaz, 760
Salem, Nourah, 519
Samadani, Ali, 729
Sanchez Carmona, Vicente Ivan, 392
Santoro, Mario, 595
Sarkar, Suparnojit, 1041
Scheuermann, Richard, 718
Searle, Thomas, 191, 665, 992
Sedoc, João, 1014
Shah, Vishwaa, 326
Sharma, Arjun, 282
Shepard, Alex, 342
Sikder, Mustafa, 128
Singh, Ujjwal, 282
Sinha, Aman, 413, 687, 864
Slavin, Rocky, 890
Smalheiser, Neil, 770
Soley, Nidhi, 1014
Son, Young-Han, 695
Sonbol, Mohamad, 944
Soni, Sarvesh, 353, 481
Stachura, Damian, 955
Stevenson, Mark, 843
Sun, Kai, 605
Sun, Yunhao, 250
Supranes, Michael Van, 997
Sutton, Adam, 191, 992

Sylvestre, Julien, 97

Takamura, Hiroya, 312, 407
Tamura, Akihiro, 679
Taneja, Sanya, 729
Tang, Buzhou, 1101
Tanwar, Ashwani, 605
Tao, Guanyu, 605
Taylor, Casey, 1014
Temprosa, Chelsea Dominique, 1029
Teo, James, 20
Tian, Shubo, 475, 502
Tizpaz-Niari, Saeid, 490
Torous, John, 932
Tran, Thao, 435
Tsujimura, Tomoki, 407
Turner, Dan, 292

Umeda, Shunnosuke, 679

Vakil, Nidhi, 760
Verstraete, Hans, 729
Veseli, Blerta, 426

Waghmare, Mayank, 104
Wakamiya, Shoko, 997
Wang, Siyao, 605
Wang, Wei, 932
Wang, Xing David, 75
Wang, Yichen, 679
Wang, Yifan, 426
Wang, Zimu, 932
Wei, Renxiong, 1101
White, Elizabeth, 519
Wilbur, W John, 475, 502
Winston, Joel, 20
Wong, Adrian, 760
Wu, Keying, 1101
Wu, Yuyang, 268

Xiang, Jiayi, 1101
Xie, Natalie, 475
Xie, Qianqian, 382, 1101
Xu, Hua, 156
Xu, Justin, 115, 426
Xu, Yihuai, 1101
Xue, Qiyao, 268
Xue, Yong, 605

Yan, Zhiyuan, 1056
Yang, Kailai, 932
Yang, Ke, 890
Yano, Ken, 312
Ye, Yanqing, 1101
Yeganova, Lana, 475, 502
Yoo, Hyunwoo, 141
Yoo, Jaehyo, 572

Zaidi, Sami, 342
Zecevic, Agathe, 181
Zeki, Sebastian, 181
Zeng, Cheng, 1101
Zhang, Jingqing, 605
Zhang, Kaidi, 1056
Zhou, Yujia, 156
Zirikly, Ayah, 1014